UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IMPROVING THE EnKF FOR HISTORY MATCHING: MULTISCALE PARAMETERIZATION AND BOOTSTRAP-BASED SCREENING

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

YANFEN ZHANG
Norman, Oklahoma
2010

IMPROVING THE EnKF FOR HISTORY MATCHING: MULTISCALE PARAMETERIZATION AND BOOTSTRAP-BASED SCREENING

A DISSERTATION APPROVED FOR THE
MEWBOURNE SCHOOL OF PETROLEUM AND GEOLOGICAL ENGINEERING

BY

Dr. Dean S. Oliver, Chair

Dr. Faruk Civan

Dr. I. Yucel Akkutlu

Dr. Deepak Devegowda

Dr. Ming Xue

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who have been very helpful to me and deserve deep appreciation.

First of all, I would like to thank my advisor Dr. Dean S. Oliver. It is difficult to overstate my gratitude to him. He is a humorous, patient, and knowledgeable person. His passion and carefulness for research deeply impressed me. He has a profound impact on my outlook and altitudes towards research. He provided me with wise advises, a great deal of freedom to explore my research interest, opportunities of involving in wide research community and petroleum industry. Special thanks goes to Mary Oliver for her caring and preparation of delicious food and fun games on every Thanksgiving.

I would like to express my gratitude to Dr. Faruk Civan, Dr. I. Yucel Akkutlu, Dr. Deepak Devegowda and Dr. Ming Xue for serving on my Ph. D. committee and for their encouragement and insightful comments on my work. I am also grateful to Dr. Younane Abousleiman for his support. I would also like to extend my thanks to all the other faculty and staff of the Mewbourne School of Petroleum & Geological Engineering.

I would like to acknowledge the financial support from the OU Consortium on Ensemble Methods (OUCEM) during my Ph.D. studies at OU. I would also like to acknowledge the donation of multiple licenses of ECLIPSE by Schlumberger, and the computational resources provided by the OU Supercomputing Center for Education and Research (OSCER). Special thanks goes to Dr. Henry Neeman (the director of OSCER) as well as the support staff for their generous help on solving supercomputer related issues. I am also grateful to Chevron for providing me the deepwater reservoir

model PFJ2 for testing, and for offering me with the internship opportunities for two summers. I had a great time as an intern at Chevron ETC and enjoyed working with people from different disciplines.

I also would like to take this opportunity to extend my appreciation to three my beloved teachers who had significant influence on me: Sen Fan, Yanchun Zhang and Changhui Chen. I want to thank all my colleagues from the OUCEM research group, especially Ning, Yan, Yaqing, Hemant and Yao. Thanks also goes to my dear friends (not listed here, but stay in ♡), without whom my life would have been pale.

Lastly and most importantly, I wish to thank my family in China for their unconditional love. To my parents, I dedicate this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Although the ensemble Kalman filter (EnKF) has been remarkably successful for history matching and quantifying uncertainty in petroleum reservoirs, there have been problems with the use of small ensembles, and occasionally with applications to real reservoirs.

Geological complexity and limited access to the subsurface typically result in a large uncertainty in reservoir properties and forecasts. There is, however, a systematic tendency to underestimate such uncertainty, especially when rock properties are modeled using Gaussian random fields. In this dissertation, the uncertainties in multiscale reservoir parameters are quantified through a stochastic multiscale model. The multiscale parameters including regional trend coefficients and heterogeneities can be estimated using the EnKF for history matching.

The proposed method of EnKF with multiscale parameterization was tested on a deepwater field whose reservoir model has over 200,000 unknown parameters. The match of reservoir simulator forecasts to real field data using a standard application of EnKF had not been entirely satisfactory, as it was difficult to match water cut in a main producer of PFJ2 reservoir. None of the realizations of the reservoir exhibited water breakthrough using the standard method. By adding uncertainty in large scale trends of reservoir properties, the ability to match the water cut and other production data was improved substantially. The results indicate that an improvement in the generation of the initial ensemble and in the variables describing the property fields give an improved history match with plausible geology. The multiscale parameterization of property fields reduces the tendency to underestimate uncertainty while still providing reservoir models that match data.

Another aspect that this dissertation focuses on is statistical sampling error caused by a limited ensemble size. The number of ensemble members (ensemble size) is critical to the efficiency and performance of the EnKF. When the ensemble size is small, the Kalman gain generally can not be well estimated. The most common approach for reducing the effect of spurious correlations on model updates is multiplication of the estimated covariance or Kalman gain by a tapering function that eliminates all correlations beyond a prespecified distance. The distance-based localization, however, is not always appropriate. In the dissertation, we propose a more general method for regularizing Kalman gain, which discriminates between the real and the spurious correlations in the Kalman gain matrix by using the bootstrap resampling technique to assess the confidence level of each element from the Kalman gain matrix.

Both the bootstrap-based screening and the commonly used distance-based localization are type of regularization methods, but use different mechanisms. The concept of distance-based localization was originally applied to the covariance matrix. Improved results, however, were also obtained by applying localization on the Kalman gain. In spite of the widespread applications of these two ways of using localization, little in the literature addresses the difference between these two ways of applying localization. This dissertation presents a comparison between the covariance regularization and the Kalman gain regularization in three aspects: improvement observed in the estimates of the Kalman gain, quality of data prediction, and the estimates of model variables. Two regularization methods are taken into consideration including the distance-dependent localization and the bootstrap-based screening. The investigation resulted in three primary conclusions. First, if localizations of two covariance matrices are not consistent, the estimate of the Kalman gain will generally be poor at the observation location. The consistency condition can be difficult to apply for nonlocal observations. Second, the estimate of the Kalman gain that results from covariance regularization is generally subject to greater errors than the estimate of

the Kalman gain that results from Kalman gain regularization. Third, in terms of removing spurious correlations in the estimation of spatially correlated variables, the performance of screening Kalman gain is comparable as the performance of localization methods (applied on either covariance or Kalman gain), but screening Kalman gain outperforms the localization methods in terms of generality for application, as the screening method can be used for estimating both spatially correlated and uncorrelated variables, and moreover, no assumption about the prior covariance is required for screening method.

CHAPTER I

INTRODUCTION

History matching is a process of adjusting reservoir simulation model parameters to match measurements. In history matching, direct measurements such as core data provide the information of reservoir properties around wellbores and can be used for building geological reservoir simulation model and generating conditional realizations of rock property field, however, such data are not always available. Indirect measurements of reservoir properties are usually obtained at the surface, in the form of seismic data and production data (such as fluid rates and well-head pressure etc.). The production data are sensitive to a large area of reservoir properties and can be obtained with a relatively low cost as the use of permanent downhole gages and surface equipment. As a result, most of the time, reservoir properties are estimated through matching production data.

In the past, history matching was done manually based on reservoir engineers' knowledge and experience. The traditional manual history matching method, currently, is still used for doing local adjustment. Manual history matching involves adjusting model parameters according to the physical relationship between data and parameters. Such manually trial and error approaches, however, are impractical to achieve well by well history match for most of the medium or large history matching problems involving a large number of wells and complicated geological settings. Hence, a lot of research is being done in assisted history matching methods. Oliver and Chen (2010) review recent developments in reservoir history matching area. Experimental design is an often used history matching method in the petroleum industry. The power of experimental design lies in identifying the most significant parameters,

the number of parameters that can be estimated, however, is very limited. Evolutionary algorithms (such as genetic method) are global optimization method, but the efficiency of such algorithms are very low. Gradient-based methods are usually effective at finding good estimates of model parameters, but calculating gradients is expensive and dependent on the source code of simulator. As the growth in the application of optimization in reservoir management, greater demands are placed on the application of history matching. The history matched models should not only reproduce the historical production behavior, but also preserve geological realism and quantify forecast uncertainty. To meet these challenging requirements on history matching method, more effective and efficient techniques are needed. The most prominent technique that is receiving growing attention in petroleum science is the ensemble Kalman filter.

1.1 The ensemble Kalman filter for history matching

The ensemble Kalman filter (EnKF) (Evensen, 1994, 2006) is a sample-based version of the Kalman filter, which uses an ensemble of model states to provide uncertainty quantification in reservoir characterization and production predictions. The application of EnKF in petroleum engineering started recently (Lorentzen et al., 2001; Nævdal et al., 2002). Since then, the EnKF was successfully applied on many synthetic cases for solving more complicated data assimilation problem (Brouwer et al., 2004; Gu and Oliver, 2005; Nævdal et al., 2006). Promising results were also obtained from several applications of the EnKF on history matching multiphase flow models of real fields (Skjervheim et al., 2007; Evensen et al., 2007; Bianco et al., 2007; Haugen et al., 2008; Zhang and Oliver, 2009). The EnKF has been shown to be a viable method for tackling the challenging history matching and uncertainty quantification problem. Many features make EnKF attractive and useful in many different contexts.

First, it is relatively easy to adapt EnKF to a number of different types of model parameters. Initially, EnKF was used to estimate spatially correlated rock properties such as porosity and permeability, but the types of variables that could be estimated using the EnKF has been expanded dramatically. The estimation range has reached to discontinuous reservoir rock facies (Liu and Oliver, 2005b,a; Agbalaka and Oliver, 2008, 2009), vertical transmissibility multipliers and fault transmissibilities (Evensen et al., 2007), relative permeability curves (Chen and Oliver, 2010), fluid contacts (Evensen et al., 2007; Thulin et al., 2007), geomechanic properties such as stress, strain, and displacement fields (Chang et al., 2010), and reservoir structural parameters (Seiler et al., 2009).

Second, the ensemble itself represents the uncertainty in reservoir characterization and production predictions (Gao et al., 2006; Zafari and Reynolds, 2007). The correlations between model variables and data are directly estimated from the ensemble of model variables and their corresponding simulated data. There is no need to compute derivatives of data with respect to changes in model variables, and computing the full covariance is also avoided by only computing the parts of the covariance that is required for the Kalman gain. These pieces have much lower dimension than the full covariance. The reservoir simulator used for modeling reservoir fluid flow and reporting the simulated data, is treated as a black box in the EnKF framework, thus no adjoint codes are required. The EnKF method can be coupled with any existing reservoir simulator.

Third, observations are sequentially assimilated whenever they are available. Thus, the ensemble Kalman filter can be used for real-time data assimilation, reservoir monitoring, and performance prediction (Nævdal et al., 2003, 2005), all of which are required for closed-loop reservoir management (combining history matching with production optimization) (Lorentzen et al., 2006; Wang et al., 2007; Chen et al., 2009).

Any method has its limitation and application requirements. A successful application of EnKF is based on the assumptions that prior distribution of model variables is approximately Gaussian and the nonlinearity of the relationship between data and model variables is not very strong, and that the ensemble size is sufficiently large. These assumptions present challenges for applying the EnKF on highly non-linear and non-Gaissian problems or problems involving high-dimensional model. Oliver et al. (2010) summarizes some of the key points of the data assimilation problem for multiphase flow in petroleum reservoirs that make the problem distinctly different from data assimilation problems in other areas such as weather and oceanography. A recent comprehensive review by Aanonsen et al. (2009) addresses various challenges for the application of the EnKF in petroleum science. This dissertation focuses on some of the challenging problems, particularly related to non-Gaussian prior distributions of model variables, implementation of the EnKF on large-scale oil fields, and statistical noise in the estimates of covariance and Kalman gain.

1.2 History matching hierarchical-Gaussian rock property fields

The EnKF has two recursive steps: forecast and analysis. In the analysis step, the update equation is almost identical as that is used in the Kalman filter except that the covariance is directly computed from the ensemble and ensemble members are updated individually using the common Kalman gain (that is calculated based on the sample covariances approximated from ensemble) to represent the posterior distribution of model variables. In other words, only the first two statistical moments (mean and covariance) are used for updating in the analysis scheme of EnKF. If the prior distribution of model variables are approximately Gaussian, the first two moments are sufficient for describing the distribution, otherwise, the EnKF can result in incorrect estimates of models variables. For optimal estimation, the implicit Gaussianity

assumption of EnKF should not be violated. In reality, however, the distributions of rock properties of many oil-bearing reservoirs appear to be non-Gaussian or non-stationary due to the diverse depositional environments.

If Gaussian random field model is used to simulate such non-stationary rock property fields, the EnKF can be used to calibrate the estimates, however, the spatially varying mean of property field can not be accounted using Gassian simulation, and moreover, the uncertainty associated with the estimates will be underestimated. Therefore, characterizing the multiscale nature of reservoir properties is essential for reliable reservoir history matching and management. In the dissertation, we describe a hierarchical method for representing and updating multiple scales of heterogeneity in the ensemble Kalman filter. The proposed method of EnKF with multiscale parameterization has been tested on a fairly large deepwater reservoir (Zhang and Oliver, 2009).

1.3 Tackling sampling error in EnKF

Another focus of the dissertation is on eliminating the statistical noise or sampling error caused by a limited ensemble size. The number of reservoir parameters to be estimated is usually quite large, because the number of grid blocks in a numerical reservoir simulation model is frequently 10^5 or larger and there are often several unknowns per grid block. In petroleum engineering related applications, the forecast step in the EnKF is done by running the numerical reservoir simulator, which is usually more expensive than the matrix computations in the analysis step, especially for a model with local grid refinement or involving gas flow. It is necessary to reduce computational demands to enable the practical application of the EnKF for reservoir history matching. Therefore, it is always desirable to use a small ensemble size to avoid running a large number of flow simulations. One problem that arises as a result of a small ensemble size, however, is spurious correlations in the sample covariance

and the Kalman gain approximated from the ensemble, which can lead to unrealistic updates to the model parameters and state variables. The cumulative effect of unrealistic updates is the loss of the ensemble variability and final break down of EnKF (Lorenc, 2003). Since the ensemble Kalman filter is a type of reduced order filter, the other problem that often arises as a result of small ensemble size is rank deficiency of the sample covariance.

For reducing the negative effect of spurious correlations and improving the effective rank, a denoising process applied on either the estimates of covariance or the Kalman gain seems to be necessary. Most of the time, the model parameters to be estimated are spatially correlated variables, and a common method for reducing the harmful effect of spurious correlations is distance-dependent covariance localization. The concept of localization in the EnKF framework was first introduced by Houtekamer and Mitchell (1998), where they simply applied a distance-based cutoff to the Kalman gain. Since then, the localization method has evolved from a distance cutoff approach to a tapering form (Houtekamer and Mitchell, 2001; Hamill et al., 2001; Houtekamer et al., 2005; Furrer and Bengtsson, 2007). The implementation of covariance localization using a tapering function is achieved by performing a Schur product of tapering function and covariance matrix. The distance-based localization is typically applied to the covariance matrix. Alternatively, the distance-based localization can be implemented on the Kalman gain (Anderson, 2007; Agbalaka and Oliver, 2008; Zhang and Oliver, 2009; Chen and Oliver, 2010), which also leads to improved results. In spite of the wide use of these two ways of performing localization, little in the literature addresses the difference between these two ways of applying localization. This dissertation presents a comparison study between the covariance localization and the Kalman gain localization.

Distance-dependent localization is an effective method, but there are some challenges associated with this method. Determining an optimal range (or correlation

length) for the tapering function is not trivial. The pre-specified range parameter in the tapering function determines the distance beyond which the correlation values are set to be zero. Hamill et al. (2001) and Lorenc (2003) showed that the optimal range for a tapering function is generally related to the ensemble size. Furrer and Bengtsson (2007) derived an expression for calculating the optimal taper function based on the true covariance, but while it may be possible to estimate the true covariance prior to data assimilation, it is extremely difficult to estimate the covariance after assimilating general flow observations in a sequential data assimilation process. Chen and Oliver (2009) applied distance-based covariance localization in sequential data assimilation for multiphase flow model. The authors point out that the region of non-zero cross-covariance cannot be determined from the region of sensitivity alone but the contribution of the prior covariance must also be considered. In their study, the authors also presented several critical conclusions about the application of localization to the covariance matrix for data assimilation in reservoir fluid flow system. They concluded that different types of data may require different types of localization, for example, for well oil production rate, a region surrounding the production well is often enough, but for well water production rate, we may want to include the injection well into the localization region. The same type of data might require different localization at different times because of the dynamics of flow system. Their study also showed that different model parameters and state variables may require different localizations. The dynamic state variables (e.g. phase saturation and pressure) are generally difficult to be localized. The optimal localization depends on the history of previous data assimilation. It is clear that applying distance-dependent covariance localization in a proper way for reservoir flow data assimilation is difficult, as a lot of factors have to be taken into consideration. Moreover, distance-dependent covariance localization is only appropriate for spatially correlated variables and it is not suitable for localizing global reservoir variables such as fluid contacts and relative

permeability parameters.

Recently, more general methods without assumption of distance dependence have been proposed for solving the sampling error caused by a small ensemble size (Anderson, 2007; Hacker et al., 2007). In the hierarchical ensemble filter method of Anderson (2007), the reliability of the regression coefficients (similar to the Kalman gain) is estimated using a group of ensembles. Confidence factors are defined to indicate the reliability of estimates, which are computed from the group of ensembles. By multiplying the confidence factors to the regression coefficients, the effect of sampling errors on the updates is reduced. Although the hierarchical method provides a mechanism for discriminating between real and spurious correlations, the extra computational cost of propagating a group of ensembles limits the practicality of its application.

Motivated by the need of more flexible localization scheme with low computational cost, a bootstrap-based screening method (Zhang and Oliver, 2010b) was developed to assess the confidence level of each element from the Kalman gain matrix and filter out the unrealistic correlations from the Kalman gain. Bootstrap is a computer-based technique for making certain types of statistical inferences (Efron and Tibshirani, 1993). In particular, it can be used to provide a measure of accuracy for estimates of statistical parameters. In the context of this investigation, bootstrap is used to compute multiple replicates of the Kalman gain matrix by resampling the same ensemble with replacement. Based on the empirical distribution of bootstrap replicates, the screening factor is inferred. In the dissertation, several different ways of defining screening factors are demonstrated. We also applied the bootstrap-based screening algorithm on the covariances. A further investigation is done on evaluating the performance of applying the bootstrap-based screening on the Kalman gain and on the covariances (Zhang and Oliver, 2010a). The investigations are carried out through two examples: a 1D linear problem for which the exact solution can be computed and a 2D highly nonlinear reservoir fluid flow problem. Consistency conditions for

covariance regularization (including the distance-based localization and bootstrap-based screening) are discussed and error evolutions in the Kalman gain regularization and covariance regularization are derived.

1.4 Scope of dissertation

The main contributions of this dissertation include four parts:

- multiscale parameterization of non-Gaussian rock properties and estimation of the multiscale parameters using the ensemble Kalman filter for history matching production data (Chapter 2);
- history matching of a deepwater reservoir located in Gulf of Mexico using the EnKF with multiscale parameterization and a comparison of the results with those obtained from the standard EnKF and manual history matching (Chapter 3);
- development of bootstrap-based screening methods to reduce statistical noise in the estimates of Kalman gain and improve robustness of the estimation (Chapter 4);
- evaluation and error analysis on covariance regularization and Kalman gain regularization, for two regularization methods: distance-dependent localization and bootstrap-based screening (Chapter 5).

Finally, Chapter 6 draws the conclusions on the major findings of this dissertation, as well as suggestions for future research.

CHAPTER II

THE ENSEMBLE KALMAN FILTER WITH MULTISCALE PARAMETERIZATION

2.1 Introduction

There is an almost universal tendency for people to underestimate uncertainty, as a result of lack of complete knowledge of the factors contributing to the uncertainty. In some cases, even though the factors contributing to uncertainty are known, there is no practical way to quantify the uncertainty. Many projects end up costing much more than the initial estimates. Such cost underestimation has a long history due to lack of adequate uncertainty quantification (Capen, 1976). In the petroleum industry, advances in seismic technology have improved our ability to image the reservoir architecture, but seismic technology alone is incapable of determining flow properties and small-scale features needed for reservoir models. Moreover, different geological processes can result in significant variation in reservoir facies distribution and petrophysical properties. In addition, optimal sampling and observations are always limited (Yang et al., 2000). As a result, a high degree of uncertainty exists when building reservoir models, especially if the reservoir lies in a complex deepwater channel systems.

Stochastic modeling is useful because it provides a systematic way of quantifying uncertainty by generating multiple reservoir property models, but the problem of systematic underestimation of uncertainty remains, especially when rock properties are modeled using Gaussian random fields (Oliver et al., 2008, pages 137–140). In

[1]Much of the material in this chapter has been accepted for publication in SPE Journal.

Gaussian simulation, it is common to either assume second-order stationarity of the random field, or that the trend (or drift) is known and the heterogeneity (or residuals) are stationary (Deutsch, 2002). But the fact is that there are always trends in the geological properties, they just appear different under different measuring scales. In this work, we model the underestimated uncertainty from regional trends by introducing stochastic trend coefficients in a polynomial trend model. The multiscale parameters including trend coefficients and heterogeneities can be estimated using the ensemble Kalman filter (EnKF) for history matching. The proposed method of EnKF with multiscale parameterization is much different from another method with a similar name, the ensemble multiscale filter (EnMSF) that was proposed by Lawniczak et al. (2009). In EnMSF, the sample covariance is replaced with a multiscale tree (nodes at the adjacent scales) at each analysis step.

One benefit of the multiscale parameterization is that it allows better uncertainty quantification of the estimates of rock properties. A second benefit is that multiscale parameterization transforms the non-Gaussian rock properties to Gaussian variables that can be better estimated using the EnKF, because the EnKF analysis step performs poorly when the variables to be estimated do not have a Gaussian distribution. Reparameterization or transformation are widely used in the EnKF related applications. The logarithmic transformation is the most common transformation method, as it is almost always applied to the permeability. Chen et al. (2007) reparameterized the state vector, replacing the non-Gaussian variable (water saturation) with an approximately Gaussian variable (the time of arrival of a saturation). Other reparameterizations have been chosen to reduce the number of model variables that must be estimated. Jafarpour and McLaughlin (2008) combine EnKF with Discrete Cosine Transform (DCT) parameterization that captures large-scale features (such as channels) by keeping only a few basis functions.

2.2 Multiscale stochastic parameters

In history matching, trends are usually considered as deterministic features and heterogeneities are treated as stochastic correlated features, the concept of which is known as universal kriging in geostatistics. The universal kriging theory, however, does not allow prior knowledge of the trend (or drift) parameters. Omre (1987) proposed Bayesian kriging which allows quantifying the uncertainty associated with the estimation of geostatistical trend parameters. Applying the stochastic model of parameterization of Bayesian kriging, we develop a method for history matching the multiscale stochastic model with the ensemble Kalman filter. Although our parameterization is similar, there are three important differences between Bayesian kriging and our proposed method. First, kriging theory requires exact data, while measurement errors are accounted for when using the EnKF to estimate geostatistical parameters. Second, kriging is a linear estimator, for which the data must be linearly related to the model variables, while the EnKF is fairly robust to nonlinearity in relationships. Thus, the dynamic production data can be sequentially assimilated to update the estimation of geostatistical parameters. Finally, production data are sensitive to model variables over large regions of the reservoir. Kriging methods do not handle observations that are sensitive to properties in large regions. The results of a field study shows that the influence of production data on the estimation of geostatistical parameters is quite significant. A detailed discussion of the results is given in Chapter 3.

The multiscale stochastic model provides a way of treating both local heterogeneities (fluctuations) and trends as stochastic features for history matching. Here, the trends we are referring to are not necessarily the very large-scale trends related to long-term changes of sediment environment, since they can sometimes be determined well from geological knowledge and seismic data, but the regional trends in

permeability or porosity that may be due to diagenesis or compaction and are difficult to identify in direct measurements. For practical modeling, the conditioning well data for estimating trends are usually sparse and regional trends are only reproduced in the neighborhood of wells, so the variance or uncertainty existing in the regional trends should be considered. Thus, we propose to use a multiscale stochastic model for representing rock property fields such as porosity and permeability.

Because the variability in the multiscale method will be quite large, it is useful to transform the property field to maintain the values of physical variables within plausible limits. In this paper, we use Ω_T to denote the transformed variables, correspondingly Ω stands for reservoir properties in the real scale. The transformation is discussed in a later section. In the multiscale model, Ω_T are expressed as the sum of a relatively large-scale trend, Θ, and a small-scale heterogeneity or fluctuation, U:

$$\Omega_T = \Theta + U \, .$$

2.2.1 Polynomial trend model

In this model of uncertainty, the small scale fluctuations U might be represented as Gaussian random fields, while the large scale trends Θ in the properties might be represented through polynomial functions of position with uncertain coefficients. If, for example, we wish to include unknown means and quadratic trends in the reservoir model, then we might write,

$$\Theta = c_1 m + c_2 x + c_3 y + c_4 x^2 + c_5 y^2 + c_6 xy \, , \tag{2.1}$$

where c_1 to c_6 are random coefficients. If set coefficients c_4 to c_6 to be zero, we will end up with a linear trend. Eq. 2.2 shows how the 2-dimensional trend is formed

using a quadratic polynomial equation with six coefficients,

$$\begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_{N_g} \end{bmatrix} = \begin{bmatrix} m & \frac{i_1}{n_x} - \frac{1}{2} & \frac{j_1}{n_y} - \frac{1}{2} & (\frac{i_1}{n_x} - \frac{1}{2})^2 & (\frac{j_1}{n_y} - \frac{1}{2})^2 & (\frac{i_1}{n_x} - \frac{1}{2})(\frac{j_1}{n_y} - \frac{1}{2}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m & \frac{i_{N_g}}{n_x} - \frac{1}{2} & \frac{j_{N_g}}{n_y} - \frac{1}{2} & (\frac{i_{N_g}}{n_x} - \frac{1}{2})^2 & (\frac{j_{N_g}}{n_y} - \frac{1}{2})^2 & (\frac{i_{N_g}}{n_x} - \frac{1}{2})(\frac{j_{N_g}}{n_y} - \frac{1}{2}) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{bmatrix}, \tag{2.2}$$

where i, j are the coordinate indices, n_x is the number of grids in $x-$direction and n_y is the number of grids in $y-$direction. N_g is the number of gridblocks in one layer. Eq. 2.2 can be expressed in a concise form as

$$\Theta = \Theta_p \Theta_c \,. \tag{2.3}$$

In Eq. 2.3, Θ represents the column vector on the left hand side of Eq. 2.2, Θ_p is the matrix of deterministic trend vectors, and Θ_c is the column vector of trend coefficients. The first column of matrix Θ_p contains the transformed uniform mean m of reservoir property field (transformation is achieved using Eq. 2.5). By multiplying a random trend coefficient c_1, we can adjust the property mean among ensemble. Different values of c_1 may be applied for individual layer of a reservoir that has multiple layers exhibiting variation in the properties among different layers. Moreover, different reservoir properties should be applied with different values of c_1 because of their different scales, for example, the scale of log permeability is nearly 10 times of the scale of porosity. All the trend coefficients are generated from the normal distribution. The polynomial trend model can simulate very diverse trends by using linear and quadratic terms, but the method is quite general. In practical modeling, it may be desirable to include other types of trends, such as a dependence on depth, according to need, but the methodology is unchanged.

2.2.2 Transformation of multiscale parameters

All model and state variables to be estimated are physical properties, so the physical constraints have to be honored while doing history matching. When variability due to uncertain trends is added to the heterogeneity, the variability of the sum can be quite large, which tends to result in unphysical values. One standard approach to solve this problem is truncation of extreme values of the parameters. But by doing so, the histogram of generated reservoir properties may have two peaks at two extreme ends in addition to a peak near the mean, in which case the benefits of using the multiscale stochastic model is substantially reduced. Thus, a transformation equation that not only can maintain a physical meaning to the generated reservoir property fields, but also reduces the negative effect of truncation can be used. The formula that we use for back-transforming reservoir properties to real scale is given in Eq. 2.4,

$$\Omega = \frac{\Omega_{\max} \exp(\Omega_T) + \Omega_{\min}}{1 + \exp(\Omega_T)} , \tag{2.4}$$

where $\Omega_{\max}$ and $\Omega_{\min}$ are the maximum and minimum plausible values of reservoir property (such as porosity and log permeability). Note that in Eq. 2.4, as $\Omega_T \to -\infty$, $\Omega \to \Omega_{\min}$ and as $\Omega_T \to \infty$, $\Omega \to \Omega_{\max}$. The transformation is more clear if one instead considers Eq. 2.5, the inverse transformation of Eq. 2.4,

$$\Omega_T = \log \left[\frac{\Omega - \Omega_{\min}}{\Omega_{\max} - \Omega} \right] . \tag{2.5}$$

The multiscale stochastic parameters are estimated using the ensemble Kalman filter that has been shown repeatedly to be an effective method for data assimilation in large-scale problems, including those in petroleum engineering. In the following section, an introduction on the ensemble Kalman filter is given.

2.3 The ensemble Kalman filter for reservoir parameter estimation

The ensemble Kalman filter is a reduced-rank sequential data assimilation method. In the standard implementation of the ensemble Kalman filter, the probability density function is approximated by an ensemble of N_e state vectors,

$$Y = [y_1, y_2, \ldots, y_{N_e}] \ ,$$

where each state vector consists of the dynamic state variables v_i and/or static model parameters m_i. Dynamic state variables are function of model parameters and change with time. For petroleum inverse problem, we usually include both the static model parameters and the dynamic state variables in the state vector

$$y_i = \begin{bmatrix} m_i \\ \\ v_i \end{bmatrix} , \qquad i = 1, 2, \ldots, N_e \ .$$

Following is a brief overview of the standard parameterization, or in other words, the commonly included state and model variables in a state vector in the standard application of the EnKF.

2.3.1 Standard parameterization for the EnKF

As the reservoir is composed of reservoir fluid (oil, water and gas) and porous rock formations, the static model parameters are related to the fluid properties and rock properties. In standard parameterization, porosity (ϕ) and permeability are the two important parameters for characterizing rock properties and are usually put into the state vector. Since permeability generally obeys log-normal distribution, we usually include the log-transformed permeability ($\ln k$) into the state vector in order to honor the implicit Gaussianity assumption of the EnKF. There may be also uncertainty in other model parameters, such as, fault transmissibility, rock facies, relative permeability curves, geomechanic properties, and the depths of fluid contacts, etc.. Certainly,

it is not necessary to estimate every uncertain parameter for a given reservoir model. We only need to identify the critical parameters and quantify the uncertainty within them. Porosity and log permeability are usually the two basic parameters to be estimated using the EnKF.

An initial ensemble of porosity and log permeability realizations are usually generated using sequential Gaussian simulation. The initial realizations can be unconditional or conditional. The term conditional implies that all realizations are conditioned to the measurements of porosity or log permeability at well locations. In this case, the values of porosity or log permeability at well locations will be honored by all realizations. The prior mean and prior covariance for generating the realizations can be determined based on the geological reservoir model and other available data (analog fields, core, logs, and seismic). The prior covariance model determines the smoothness or continuity of the random fields. The correlation length used in the theoretical prior covariance model influences the number of degrees of freedom of the parameter space and the impact of measurements on the parameters.

In the sequential data assimilation framework, history matching process is not pure parameter estimation, but the combined parameter and state estimation problem. The dynamic state variables are also included into the state vector. As mentioned previously, the dynamic state variables are function of model parameters and change with time. Such parameters include, for example, grid-based pressure and phase saturations (oil saturation S_o, water saturation S_w, and gas saturation S_g). Some physical constraints on phase saturations need to be honored during the data assimilation process. The sum of all phase saturations present in a gridblock should always be equal to 1. The water saturation in a gridblock, S_w, can not be lower than the irreducible water saturation, and oil saturation, S_o, can not be below residual oil saturation in oil/water or oil/gas system.

2.3.2 Forecast and analysis

The ensemble Kalman filter consists of two recursive steps: one is forecast step for solution of a dynamic system to a new time point and the second step is a Bayesian update for assimilating new data. In the forecast step, the model parameters m_i remain the same,

$$m_{t+1,i} = m_{t,i} \ ,$$

while the dynamic model variables v_i are evolved from time t to time $t+1$,

$$v_{t+1,i} = f(y_{t,i}) \ .$$

The predicted data for the ith ensemble member at time $t+1$, $d^f_{t+1,i}$, are computed from the model variables by running a forward model. In petroleum application, the forward model is a numerical reservoir simulator ($g(\cdot)$),

$$d^f_{t+1,i} = g(y_{t+1,i}) \ , \qquad\qquad i = 1, 2, \ldots, N_e \ .$$

The next step is analysis or update. All algorithms for the analysis step are based on the idea that the pdf of the variables, and specifically, summary statistics of the probability density function, can be estimated from an ensemble of random realizations. As data are assimilated, all realizations are adjusted using Eq. 2.6 for consistency with the new observation. In the analysis step, both the forecast model parameters and state variables are updated,

$$Y^a = Y^f + K_e(d_{obs} - d^f) \ . \tag{2.6}$$

In this expression, the subscript of time index is neglected, K_e is the Kalman gain, and d_{obs} denotes perturbed observations obtained by adding zero-mean noise with covariance C_D to the actual measurement values. The ensemble Kalman filter method includes two sources of sampling error in the analysis step: the random sampling of the initial ensemble of model realizations and random sampling in the measurement

perturbations. Although the original implementation of the ensemble Kalman filter method (Evensen, 1994) did not include perturbations to the observations, Burgers et al. (1998) and Houtekamer and Mitchell (1998) showed that by perturbing the observations, it is possible to obtain the correct variance in linear data assimilation problems with large ensembles. Others have pointed out that the addition of perturbations to the observations introduces additional sampling error, and advocate methods that do not require perturbations (Tippett et al., 2003). These methods compensate for the lack of variability by modifying the formula for updating model and state perturbations from the mean. For a linear assimilation problem, Whitaker and Hamill (2002) showed that an ensemble square-root filter (EnSRF), was better at estimating ensemble variances than the ensemble Kalman filter. Sakov and Oke (2008) showed similar benefits of mean-preserving square root filters over the standard EnKF when applied to the linear advection model of Evensen (2004). Zhang et al. (2010), however, showed that the sampling error in the observation perturbations can be largely eliminated by using second-order-exact sampling of observation perturbations and eliminating the cross-covariance between the observation perturbations and the deviations of predicted data. In this case, the performance of the EnKF and the EnSRF are nearly identical. As the problems became more nonlinear, the advantage of the EnSRF disappeared and all filters performed similarly.

The Kalman gain, K_e, is computed from the forecast ensemble Y^f and predicted data d^f, using the expression,

$$K_e = C_{yd}^f (C_{dd}^f + C_D)^{-1} \ , \tag{2.7}$$

where C_{yd}^f is the sample covariance between the variables in the state vector and predicted data; C_{dd}^f is the sample covariance between different predicted data. C_{yd}^f and C_{dd}^f are directly estimated from the ensemble.

Data assimilation for multiphase flow in porous media is particularly difficult, however, because the relationships between model variables (e.g. permeability and

porosity) and observations (e.g. water cut and gas-oil ratio) are highly nonlinear. Because of the linear approximation in the update step and the use of limited number of realizations in an ensemble, the ensemble Kalman filter has a tendency to systematically underestimate the variance of the model variables.

2.4 EnKF with multiscale stochastic parameters

The EnKF with multiscale parameterization is achieved by including the multiscale parameters into the state vector to replace porosity and log permeability that are usually used in the standard parameterization of EnKF. Thus, during the history matching process, instead of updating porosity and log permeability directly, their two different-scale components: the small-scale heterogeneity of each gridblock and the large-scale trend coefficients of each layer in a reservoir are updated. Continual updating of the large scale parameters (trend coefficients) can effectively enhance the influence of EnKF during analysis. Moreover, as mentioned before, determining trends in reservoir depends on data, but the fact is that static well data are usually limited and sparse, and that wells are preferentially drilled in regions of good reservoir properties. Thus, if only static data are used for estimating trends, there is possibility of being misled by generating reservoir property fields that are too optimistic. On the other hand, when production data with large regions of support are used for estimating trends, we may get rid of these pitfalls that occur when only static data are used. Because the procedure is slightly different from standard EnKF, we have included some details.

Application of EnKF begins with the generation of N_e unconditional realizations of U and Θ_c. The initial unconditional ensemble, Y_0^f (the subscript 0 denotes initial condition), is a collection of random initial state vectors:

$$Y_0^f = [y_{0\,1}^f, y_{0\,2}^f, \cdots, y_{0\,N_e}^f]\,.$$

Usually the relationship between porosity and log permeability is positively correlated,

therefore, the common random trend coefficients (c_2 to c_6) that control the trend shape are used for porosity and log permeability as shown in the following example of state vector j,

$$y_{0\,j}^{f} = \begin{bmatrix} U_j \\ \Theta_{c\,j} \end{bmatrix} = \begin{bmatrix} U_{\phi_1,j} \\ \vdots \\ U_{\phi_{N_g},j} \\ U_{\ln k_1,j} \\ \vdots \\ U_{\ln k_{N_g},j} \\ c_{1_\phi,j} \\ c_{1_{\ln k},j} \\ c_{2,j} \\ \vdots \\ c_{6,j} \end{bmatrix}.$$

Certainly different c_1 should be used for porosity and log permeability as the magnitudes of these two properties are very different. For some rock types, the porosity and log permeability are not very correlated, two separate sets of trends coefficients (c_1 to c_6) can be used in the state vector. If we have some prior knowledge about the reservoir, for example, well logs, an initial ensemble of conditioned realizations Y_0^a can be obtained by assimilating the static well data using Eq. 2.6. The static well data should be transformed using Eq. 2.5 and perturbed with very small noise before being assimilated.

To continually update the estimates of U and Θ_c, production data are sequentially assimilated, whenever they are available. In this sequential data assimilation process, along with U and Θ_c, the dynamic state variables such as pressure P, water saturation S_w, and gas saturation S_g (if three-phase model) are also included into the state vector. The algorithm given above is for standard EnKF in combination with the

multiscale stochastic model. A flow chart of this procedure is provided in Fig. 2.1. The multiscale stochastic model can also be easily combined with any iterative schemes of EnKF (Zafari et al., 2006; Gu and Oliver, 2007) for more effectively solving highly nonlinear problems. The primary extra computation cost of using the multiscale stochastic model is the transformation of converting multiscale parameters to reservoir properties, but such cost is negligible compared to the cost of numerical reservoir simulation.

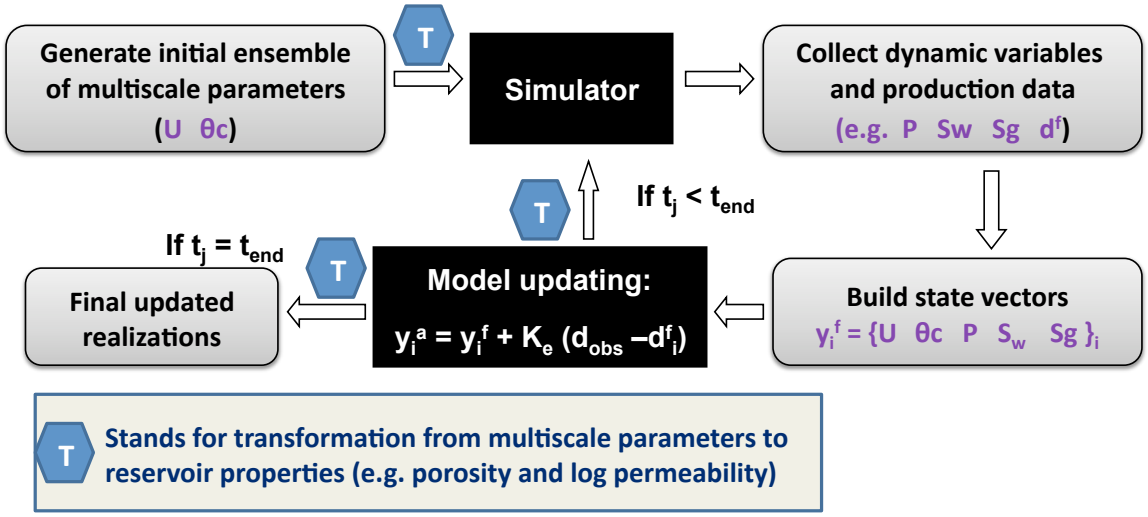


Figure 2.1: Flowchart of EnKF with multiscale parameters.

CHAPTER III

HISTORY MATCHING OF A DEEPWATER RESERVOIR

3.1 Description of field and simulation model

The deepwater reservoir PFJ2 is located in more than 1700-ft waters in the Gulf of Mexico. The reservoir rock is slightly compressible. There are 3 bottom aquifers, 2 water injection wells, and 6 production wells in the reservoir. The production wells came online at different times during the production history. Fig. 3.1 shows the structure of PFJ2. The reservoir simulation model is an Eclipse 100 three-phase black oil model (Schlumberger, 2007) with grid dimensions of 159 $\times$ 149 $\times$ 5. The total number of cells is 118,455, of which 95,379 are active. The horizontal grid has a resolution of 164 ft $\times$ 164 ft. There are 5 layers with thickness varying between 0.1 ft and 23 ft. The field has been produced for about 6 years. In the simulation model, the production wells were produced by the target oil production rate with minimum bottom hole pressure as the secondary constraint. The injectors were also on rate control with maximum bottom hole pressure as the secondary constraint.

3.2 Traditional manual history matching

Traditional manual history matching has previously been applied for this field case. Manual history matching basically is manually editing the input files of simulator according to petrophysical knowledge and experience of reservoir engineers. As observed in Fig. 3.2, manual history matching of the bottom hole pressure of producer P 5 (Fig. 3.2 (a)) was partially accomplished by decreasing the permeability within

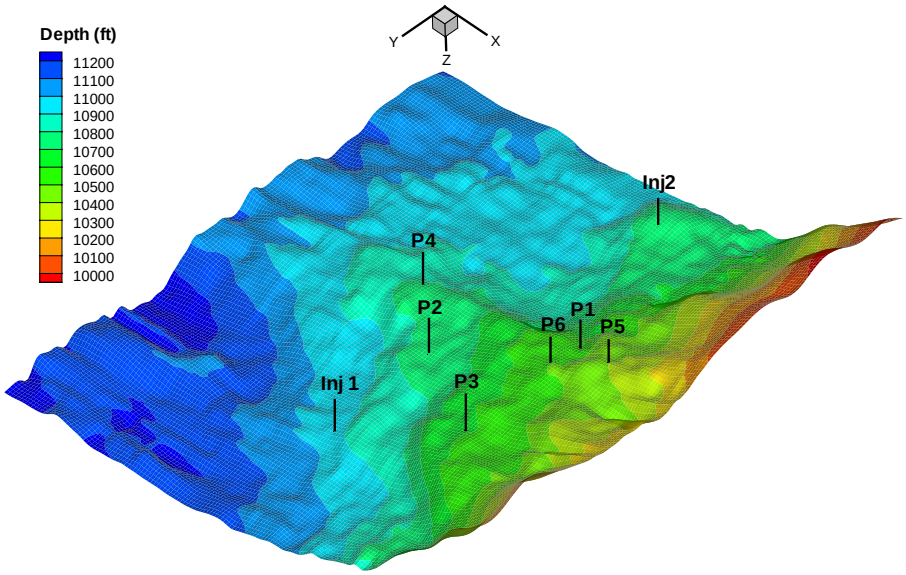


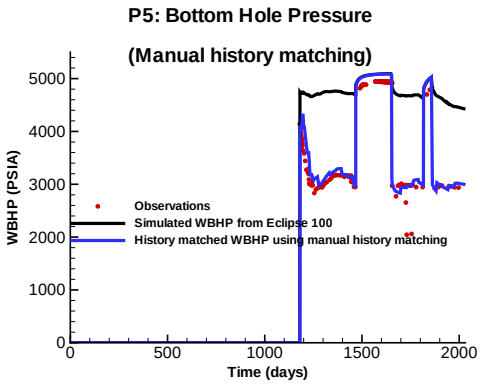
Figure 3.1: Structure map of PFJ2 field.

a zone surrounding the well, which results in a geologically unrealistic permeability field (Fig. 3.2 (b)). In order to postpone the water breakthrough of producer P 4, the pore volume of the neighborhood around well P 4 is increased by 10% (figure is not included here). On the other hand, matching the water cut history of producer P 1 (Fig. 3.2 (d)) can not be easily achieved by adjusting the pore volume nearby the well. As shown in Fig. 3.2 (c), the water front is distant from well P 1. The scattered green points in the aquifer (large blue region) in Fig. 3.2 (c) are inactive cells.

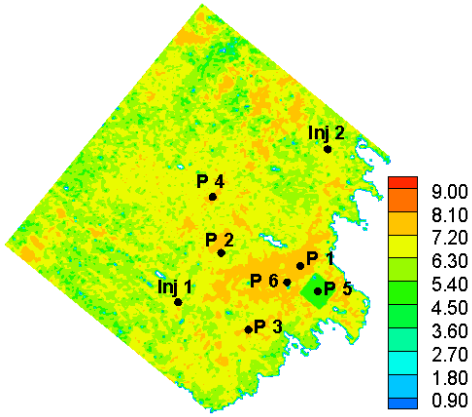
In addition, the fluctuating gas oil ratio (Fig. 3.2 (e) and (f)) data can not be easily matched using manual history matching. Manual history matching is a time-consuming trial-and-error process, and usually difficult to achieve well by well history match. Therefore, the ensemble Kalman filter is applied to solve this real field history matching problem.

3.3 History matching using the EnKF

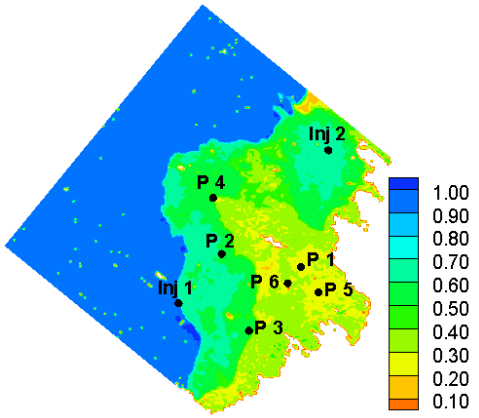
Without knowledge of data used to create the model, it is hard to identify which aspects of the model can be altered, so we have to make some reasonable assumptions



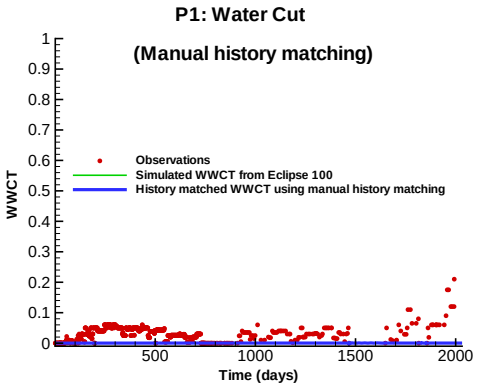
(a) P 5: Bottom hole pressure



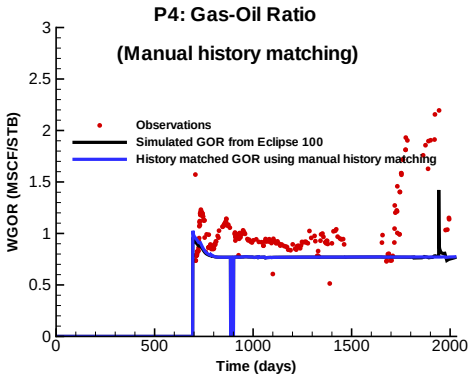
(b) History matched permeability of model layer 1



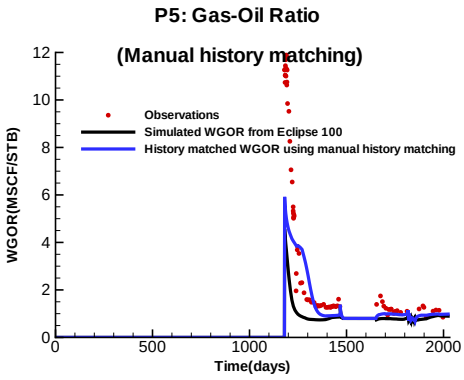
(c) Final water saturation of model layer 1



(d) P 1: Water cut



(e) P 4: Gas-oil ratio



(f) P 5: Gas-oil ratio

Figure 3.2: Traditional manual history matching results. (In the line plots, red dots are observations, black curve denotes the output from the simulation model without history matching, blue curve denotes the output from the manually history matched simulation model.)

for applying the EnKF. We have assumed that there is no uncertainty in the structural model (e.g. the locations of aquifers etc.), fluid contacts, and the relative permeability curves. We also assumed that the values of porosity and permeability at the well locations in the reservoir simulation model provided to us are sufficiently accurate to be used as static well data. In this history matching problem, we evaluated the performance of the EnKF on three cases.

3.3.1 Case 1: the EnKF with Gaussian simulation for generating initial ensemble

The standard EnKF was first tested to solve this history matching problem. The initial ensemble of porosity and log permeability was generated using sequential Gaussian simulation. The input parameters of sequential Gaussian simulation, such as mean and variogram model are obtained by carrying out statistical analysis of the given simulation model assuming ergodicity. The initial realizations of porosity are generated using a nested isotropic variogram model (0.65 Sph (4.5) + 0.35 Exp (65)), mean of 0.1, and standard deviation of 0.03. The realizations of log permeability are co-simulated using the same variogram model, mean of 2.7 and standard deviation of 1.8. The correlation coefficient between porosity and log permeability is 0.6.

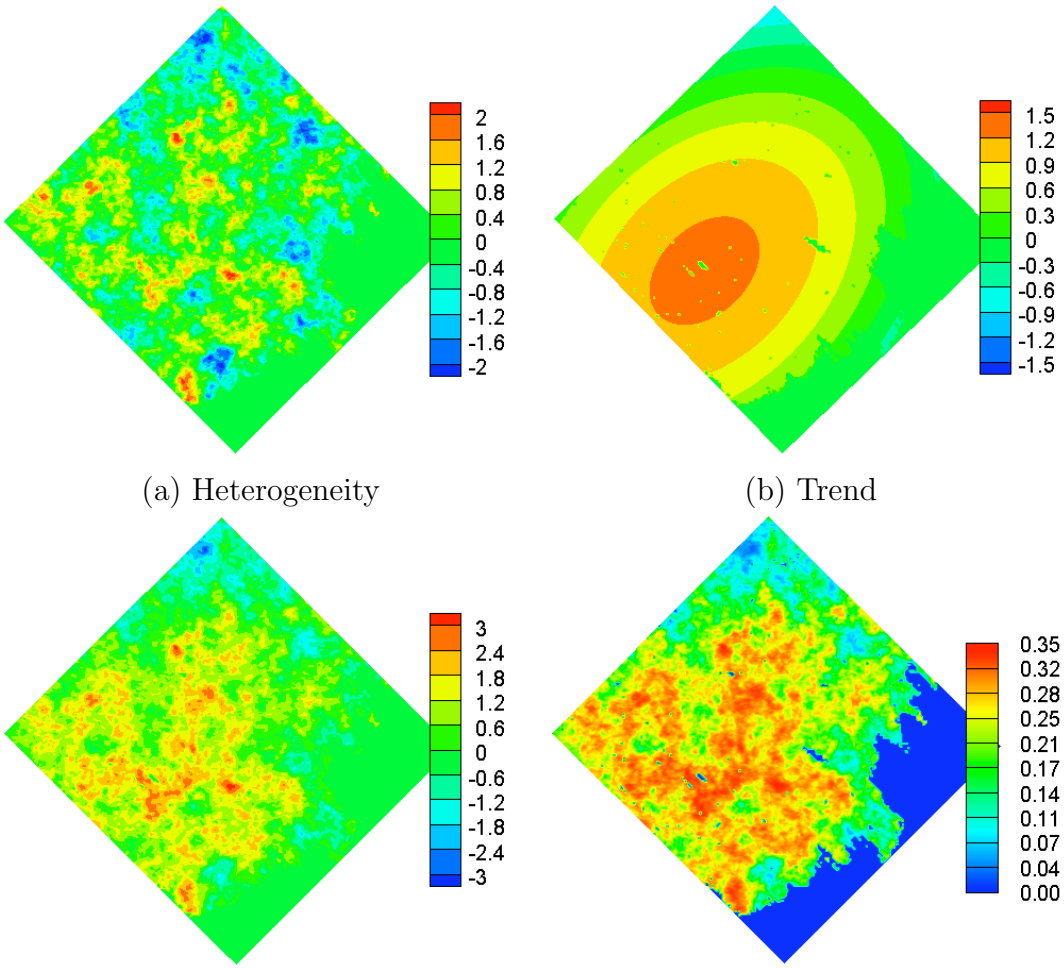
3.3.2 Case 2: the EnKF with multiscale simulation for generating initial ensemble

In this case, the multiscale stochastic model was used for the generation of the initial ensemble. The trend model used in this field study is a quadratic polynomial formulation of three terms shown in Eq. 3.1.

$$\Theta = c_1(m + m_w) + c_2 a_x (x - O_x)^2 + c_3 a_y (y - O_y)^2 \,, \tag{3.1}$$

where c_1, as mentioned previously, is sampled from a normal distribution with mean 1, m is the transformed mean of reservoir properties, and m_w is a weighting parameter for adjusting the mean when c_2 and c_3 are sampled from a normal distribution

with a nonzero mean. a_x and a_y are shape adjusting parameters, and O_x and O_y are the parameters controlling the center of trends. All the parameters shown in Eq. 3.1 are fixed for all initial realizations, except the trend coefficients (c_1, c_2, c_3) that vary with each realization in order to quantify the uncertainty in regional trends. One realization of the heterogeneity and trend used for porosity (Fig. 3.3) and those for log permeability (Fig. 3.4) shows the multiscale features achieved by using stochastic heterogeneity and trend coefficients, and also demonstrates the necessity of the transformation step.

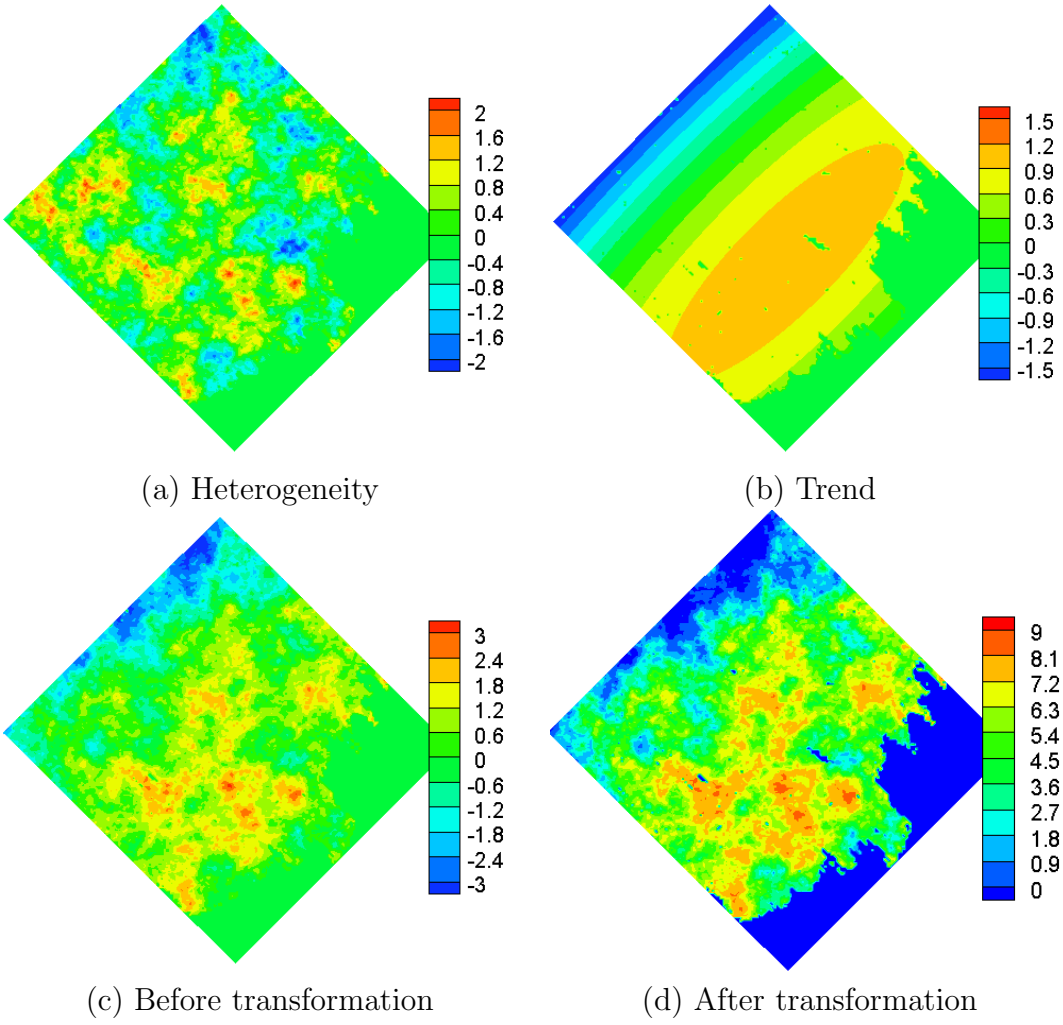


(a) Heterogeneity (b) Trend

(c) Before transformation (d) After transformation

Figure 3.3: Illustration of heterogeneities, trends and the resulting porosity (before/after being transformed).

We compare the standard deviation maps of initial ensemble from Gassian simulation (Case 1) and multiscale simulation (Case 2) in Fig. 3.5. It is evident that the standard deviations with multiscale simulation are much higher. In the models generated from multiscale simulation, the standard deviation increases significantly as we move away from the drilled region.



(a) Heterogeneity (b) Trend

(c) Before transformation (d) After transformation

Figure 3.4: Illustration of heterogeneities, trends and the resulting log permeability (before/after being transformed).

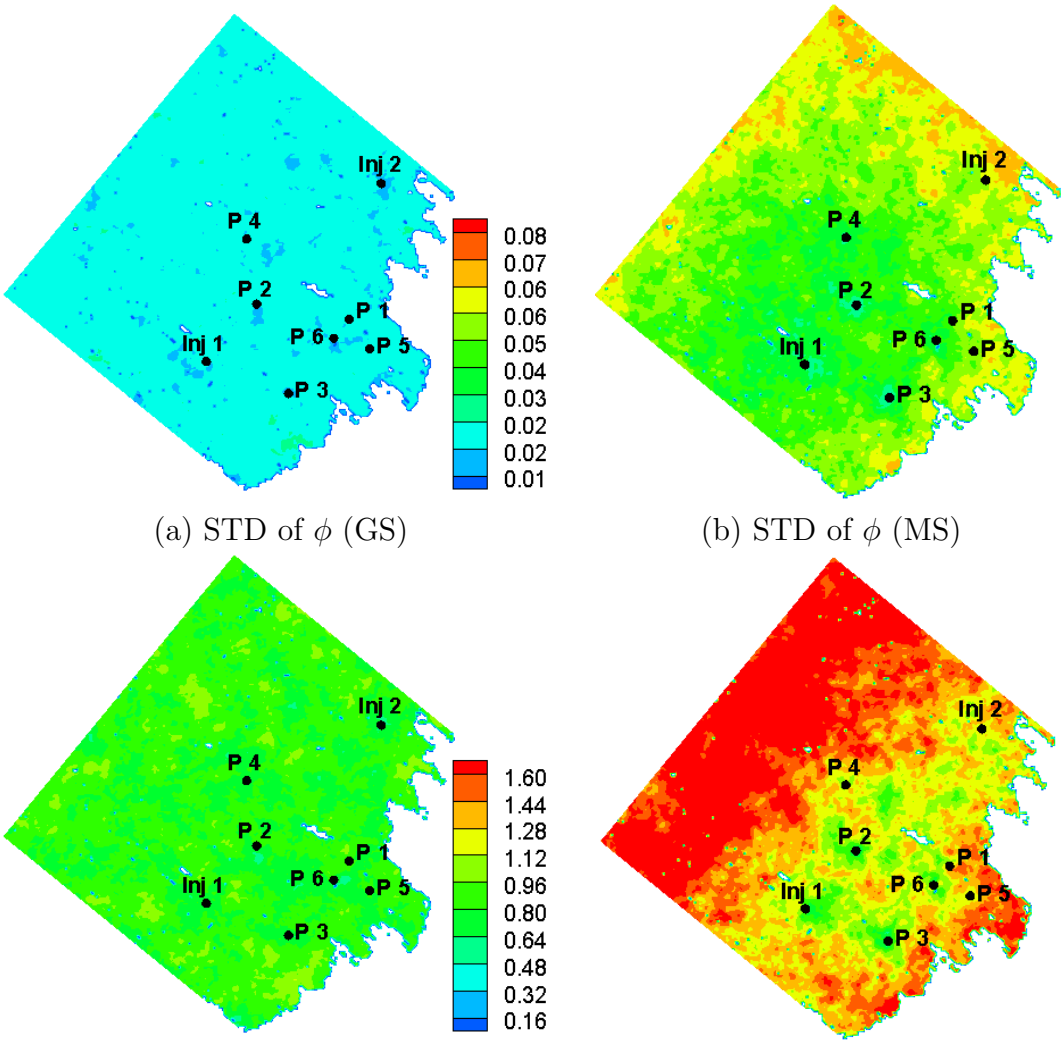
In order to examine the uncertainty covered by the initial ensemble, the initial realizations were run forward in time from day 0 to day 3000, without assimilating any production data. In fact, the production history ends at day 2032, but in order to see the forecast of late water breakthrough, simulated history was elongated to

3000 days. Fig. 3.6 shows the comparison of the predicted production based on the initial ensemble generated through Gaussian simulation and those obtained from the multiscale simulation. Water breakthrough of P 1 during the production history was observed using multiscale simulation, whereas the water breakthrough time predicted by the standard EnKF was far away from the actual observations. If the initial ensemble adequately captures the uncertainty, we should expect that the actual data fall within the range of outcomes from the ensemble. Through comparison, the spread of the realizations generated with the multiscale model gives a better representation of the initial uncertainty in the model forecasts.

3.3.3 Case 3: the EnKF with multiscale parameterization

Instead of limiting the use of multiscale stochastic model to generating the initial rock properties, we include the multiscale parameters into the state vector, replacing the porosity and log permeability as stated in Chapter 2, updating the multiscale parameters by continually assimilating the production data.

For all the three cases, the ensemble size is 60. Besides porosity and log permeability (for Case 1 and Case 2) or multiscale parameters (for case 3), the state vector also includes three types of dynamic state variables per gridblock: pressure, water saturation, and gas saturation. Thus, the dimension of a state vector is nearly 0.5 million. We implemented a parallel version of data assimilation with multi-processors. A distance-based localization scheme was used in these cases to reduce the spurious correlations and to increase the effective rank of the ensemble (Chen and Oliver, 2010; Gaspari and Cohn, 1999). The data assimilation is carried out approximately every three months and the total number of data assimilation times is 26. The assimilated production data include bottom hole pressure (well constraint), water cut, gas oil ratio, and flow rate target (well control).



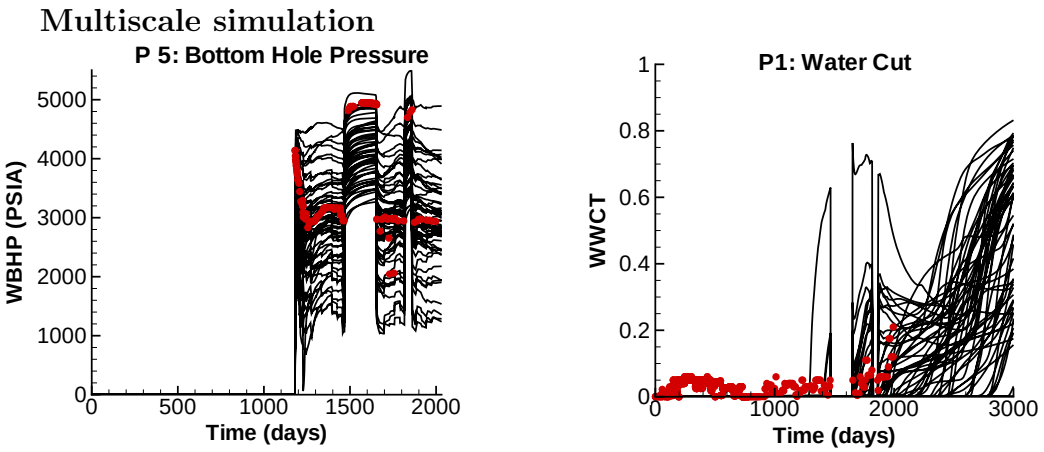
(a) STD of ϕ (GS)

(b) STD of ϕ (MS)

(c) STD of $\ln k$ (GS)

(d) STD of $\ln k$ (MS)

Figure 3.5: Standard deviation of porosity and log permeability of model layer 1. ϕ stands for porosity and $\ln k$ stands for log permeability. GS denotes Gaussian simulation, and MS denotes multiscale simulation.



P 5: Bottom hole pressure | P 1: Water cut

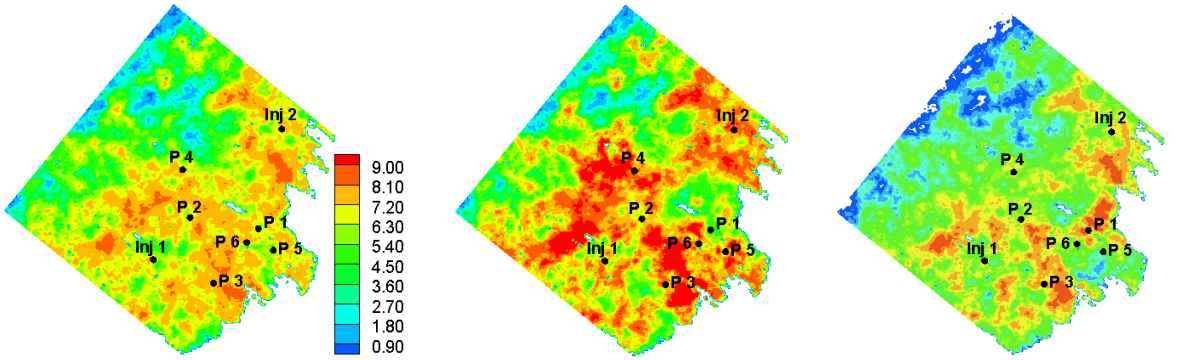
P 5: Bottom hole pressure | P 1: Water cut

Figure 3.6: The production forecast based on the initial ensembles of porosity and permeability (black lines), and the observations (red dots).

3.4 Results and discussion

In Case 2, the initial realizations of log permeability and porosity were generated from the multiscale simulation method, so the initial variability in production predictions is fairly large. Because the standard parameterization of property fields is used in this case and the spatial variability in rock properties is very large, the direct updates to the property fields (including porosity and log permeability) are not suitably regularized, and the updated result is property fields with large over- and undershoot in values. Because of the extreme values existing in the updated reservoir property fields, the simulator was unstable, it was not possible to complete the entire data assimilation, and process stopped after the 14th data assimilation time. Fig. 3.7 compares one updated realization of log permeability after 14 assimilations of data using the EnKF with multiscale parameterization (Case 3) with the corresponding results obtained using the standard parameterization but the same initial ensemble (Case 2). The properties obtained from Case 2 are unrealistic compared to the magnitudes shown in the initial $\ln k$ and updated $\ln k$ from Case 3. The results from Case 2 indicate that improved initial realizations alone may not lead to good assimilation results, suitable parameterization (or in other words, what parameters are chosen to put in the state vector for being updated) is equally important. The porosity and log permeability that are simulated using multiscale stochastic model do not follow a Gaussian distribution, although the non-Gaussianity is not so strong as to appear as bi-modal. Results from updating of such non-Gaussian parameters using the EnKF (Case 2) is not as good as results from updating the multiscale parameters using the EnKF (Case 3), since the multiscale parameters (including heterogeneity and trend coefficients) have Gaussian distributions. Because of the incomplete data assimilation of Case 2, we only compare the results from Case 1 and Case 3 in the rest of the section.

The predictions during history matching process from Cases 1 and 3 are shown



(a) Initial $\ln k$ (b) Updated $\ln k$ (Case 2) (c) Updated $\ln k$ (Case 3)

Figure 3.7: Comparison of the updated log permeability of Realization 8 from Case 2 and Case 3 after 14 data assimilation times.

in Fig. 3.8, in which only production data that were not well matched using manual history matching are included. The standard EnKF (Case 1) did not give a totally satisfactory history match as shown in the first column of Fig. 3.8. Specifically, none of the realizations were able to correctly predict water breakthrough of P 1 by implementing the standard EnKF. Moreover, the ensemble predictions of bottom hole pressure of P 5 are much higher than the observations before the shut in period. Also, although not shown here, the ensemble predictions of field gas production total are generally lower than the observations. Many factors can influence the performance of the standard EnKF, but one of the primary reasons appears to be the relatively small variability in the initial reservoir property fields, which limits the adjusting space of EnKF. By generating the initial ensemble of porosity and permeability based on the assumption of a stationary mean defined by the given simulation model, it seems that the uncertainty existing in the model was underestimated. The EnKF with multiscale parameterization was able to match water breakthrough in well P 1 and was able to obtain a better match of bottom hole pressure in P 5. Both methods, however, achieved better results than the manual history match. Fig. 3.9 shows the final estimates of water saturation maps. Comparing the final water saturation map from the EnKF with multiscale parameterization with those maps obtained from

standard EnKF and manual history matching shown in Fig. 3.2(c), it is observed that the EnKF with multiscale parameterization method has resulted in further advance of water fronts from the injectors to producer P 1.

Fig. 3.10 and Fig. 3.11 compare the final estimates of porosity and log permeability fields and the corresponding final standard deviation maps. In both cases, the standard deviations of reservoir properties are substantially reduced around the drilled area although the multiscale method maintained larger variability in the regions far from the wells. In Fig. 3.12, we see that the larger variability is a result of increased variability between ensemble members; the realizations from the multiscale method look plausible after updating when the trend parameters are included in the updating. The EnKF with multiscale parameterization is more effective than the other methods we investigated in terms of matching data and estimating model variable distributions.

Fig. 3.13 shows the histograms of the initial and final realizations of trend coefficients. We observe a large influence of production data on the estimates of trend coefficients. The largest change is a substantial reduction in uncertainty in the trend coefficients c_2 and c_3 describing the quadratic trends in the property fields by sequentially assimilating the dynamic production data.

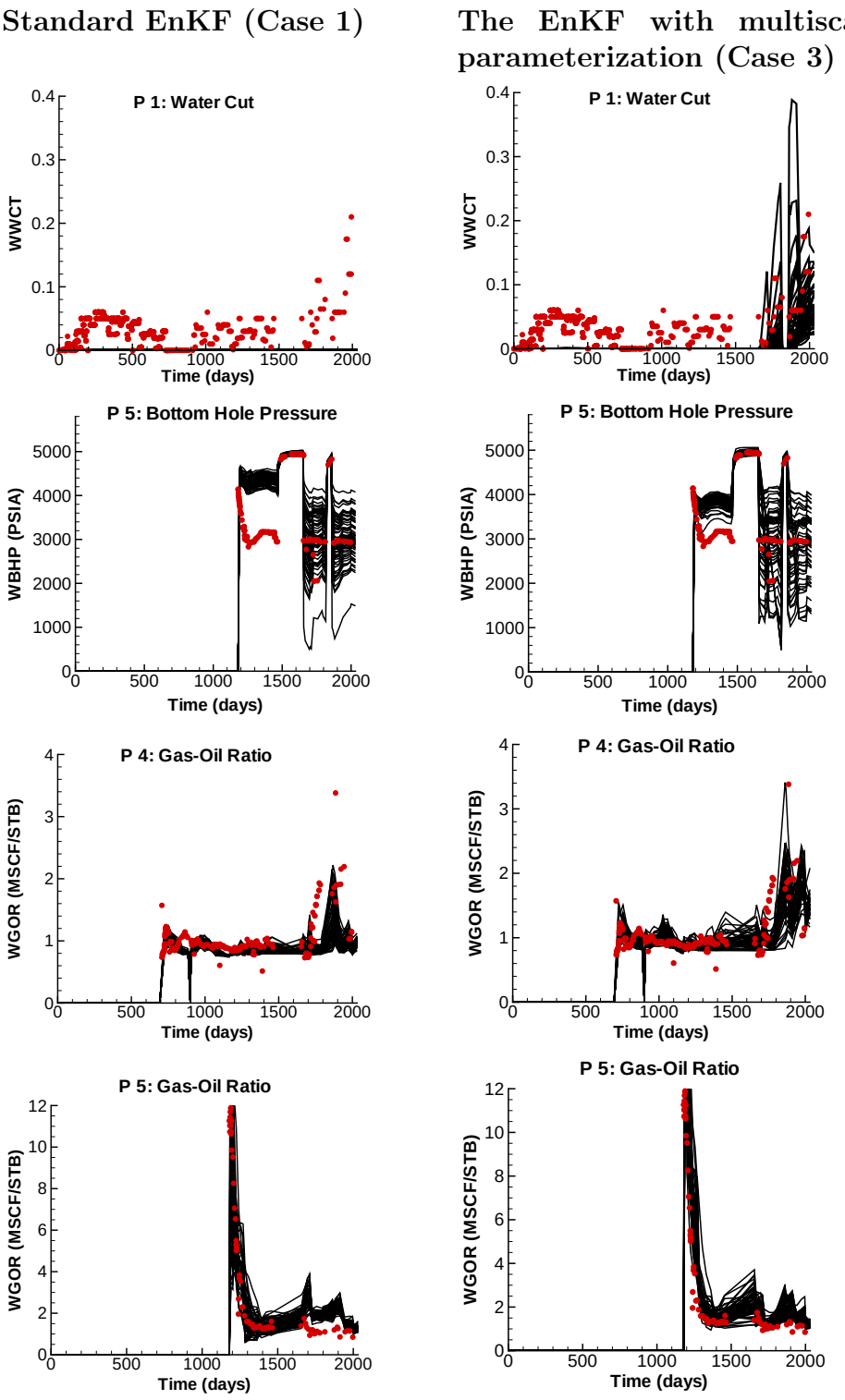
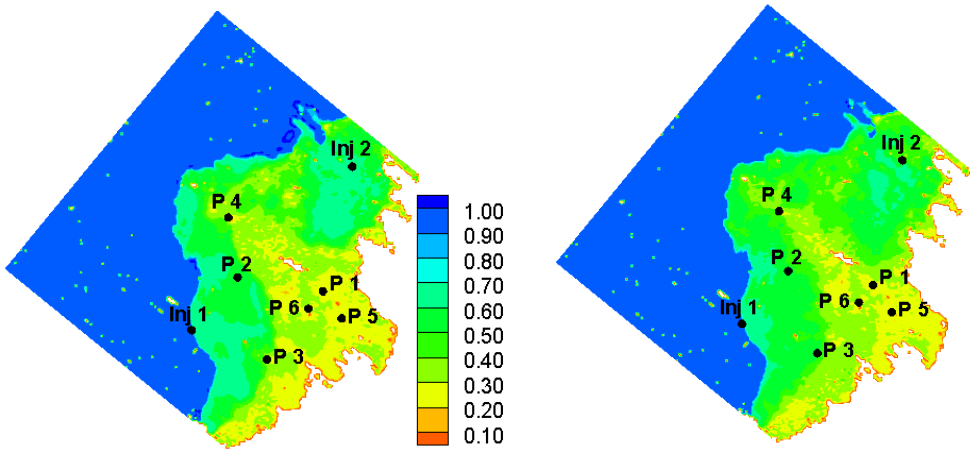


Figure 3.8: The production data during the history matching process and prediction using the EnKF with multiscale parameterization and the standard EnKF. (The black lines denote the results from different ensemble members and the red dots denote observations.)



(a) Case 1 (b) Case 3

Figure 3.9: Final estimate (ensemble mean) of water saturation of model layer 1.

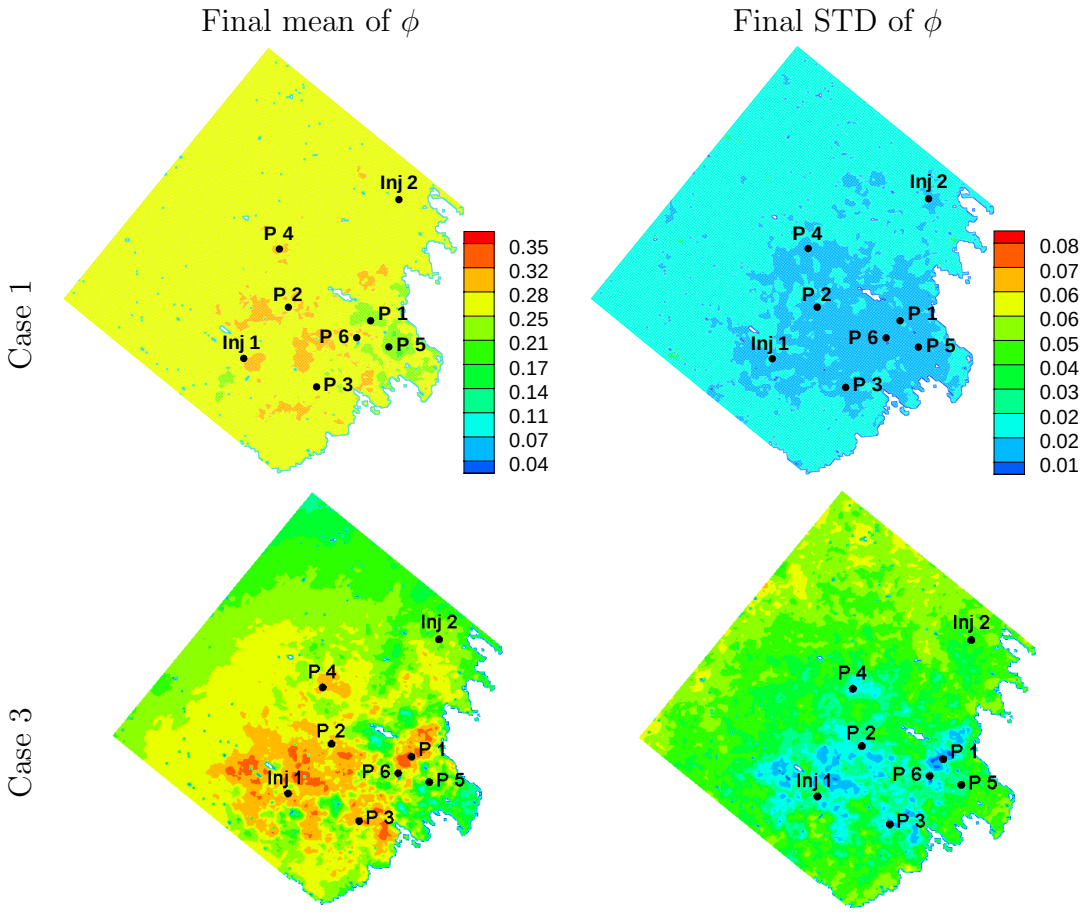


Figure 3.10: Final estimates (ensemble mean) and associated standard deviations of porosity of model layer 1.

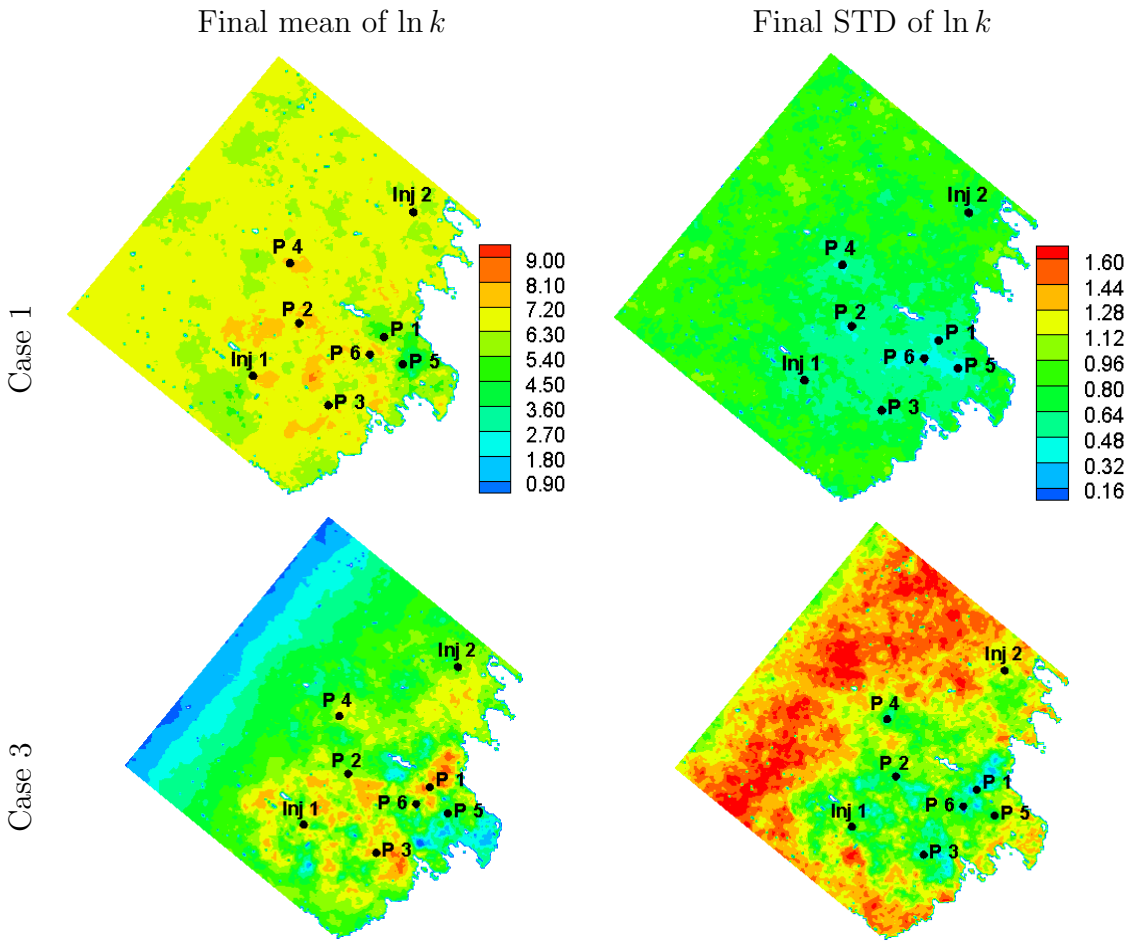


Figure 3.11: Final estimates (ensemble mean) and associated standard deviations of log permeability of model layer 1.

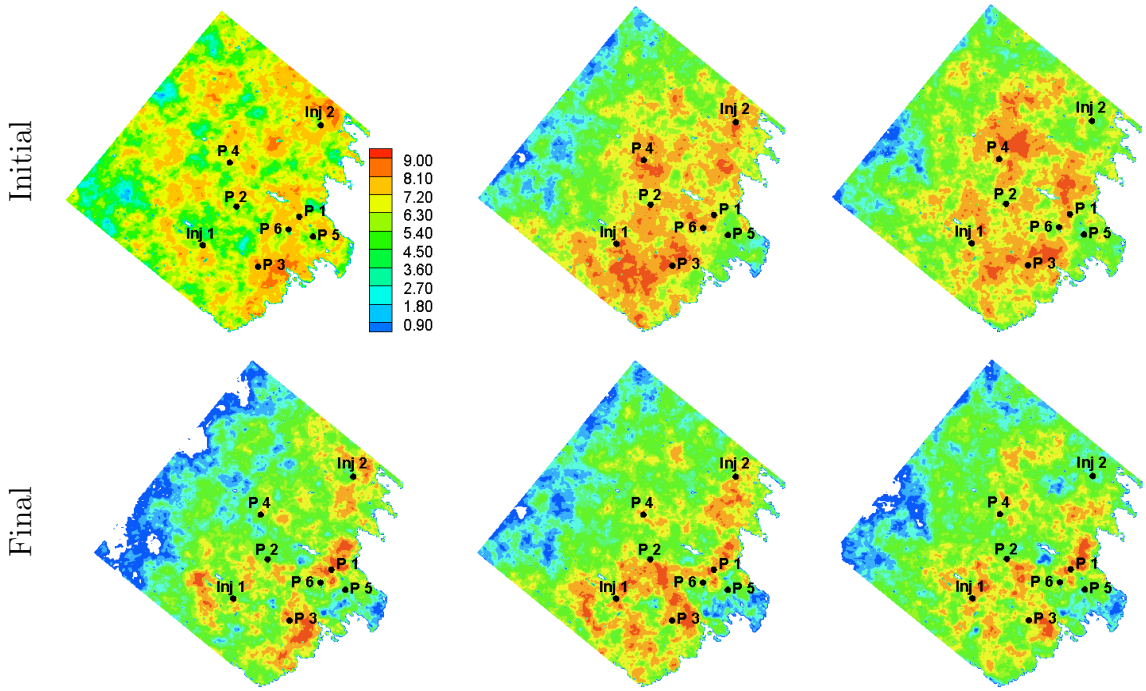
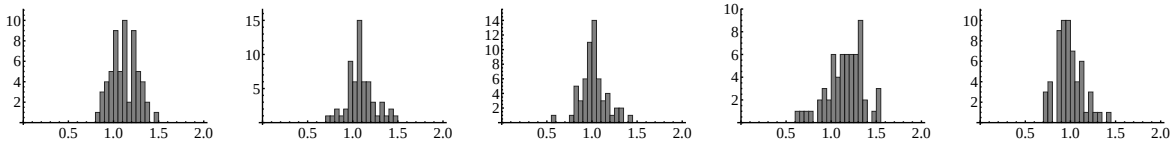


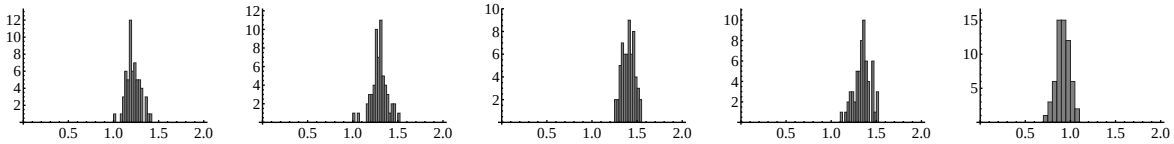
Figure 3.12: Several examples of initial and corresponding final realizations of log permeability of model layer 1 for the EnKF with multiscale parameterization (Case 3).

Initial estimate: Coefficient c_1 of porosity for different layers



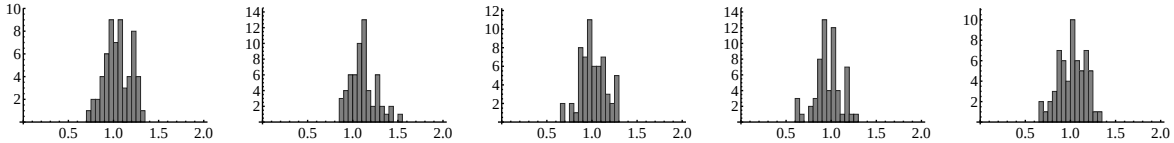
layer1 layer2 layer3 layer4 layer5

Final estimate: Coefficient c_1 of porosity for different layers



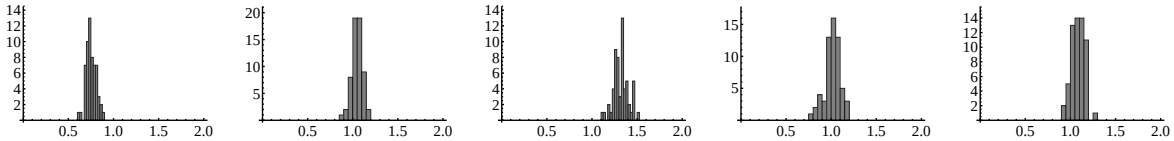
layer1 layer2 layer3 layer4 layer5

Initial estimate: Coefficient c_1 of log permeability for different layers



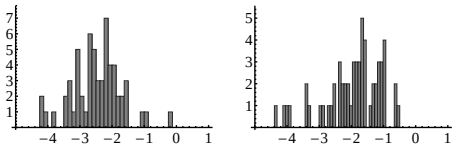
layer1 layer2 layer3 layer4 layer5

Final estimate: Coefficient c_1 of log permeability for different layers



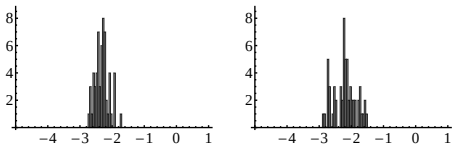
layer1 layer2 layer3 layer4 layer5

Initial estimate: Coefficients c_2 and c_3 of porosity and log permeability for all layers



c_2 c_3

Final estimate: Coefficients c_2 and c_3 of porosity and log permeability for all layers



c_2 c_3

Figure 3.13: Histograms of trend coefficients before and after assimilation of production data.

3.5 Chapter summary

Because of the complexity of the relationships between data and model variables, it is generally difficult to achieve well-by-well manual history matches that are geologically plausible. The field example discussed in this chapter shows such a case. EnKF as an assisted history matching method, circumvents these difficulties inherent to the manual process. However when EnKF is applied for history matching, because of the implicit assumption of Gaussianity, Gaussian simulation is often used for generating the initial ensemble of rock properties, which, at the same time, causes a systematic underestimation of the geostatistical uncertainty. We have shown that multiscale stochastic structure provides a way to increase the variability of reservoir property fields and avoids the overshooting problem by introducing an appropriate transformation to property fields. The results of the field case study show that the ability to match the water cut and other production data was improved by adding uncertainty in trends. Compared to the standard EnKF, the EnKF with multiscale parameterization provided better uncertainty quantification. The results also indicated that an improvement in the generation of the initial ensemble and in the variables describing the property fields gave an improved history match with plausible spatial distributions of petrophysical properties.

CHAPTER IV

BOOTSTRAP-BASED SCREENING OF KALMAN GAIN

The EnKF technique often performs well for data assimilation when the ensemble size is sufficiently large, but the computational cost grows with the ensemble size. As a result, it is always desirable to use as small an ensemble as possible. When a small ensemble size is used, however, the underlying probability distribution can not be well sampled. The introduced sampling error lead to spurious correlations in the estimates of covariances and Kalman gain. Spurious correlations are defined as the correlations that are not present in reality between two variables. The harmful effect of spurious correlations is the unrealistic changes to the model and state variables (e.g. porosity, log permeability, pressure, etc.). After several assimilation times with poor updates, it is possible that the variability in the ensemble collapses (Lorenc, 2003).

A commonly used method for eliminating spurious correlations is distance-dependent localization, however, this method is lacking in generality for application and there are several difficulties for practical application as discussed in Chapter 1. Anderson (2007) proposed a hierarchical filter, which is a general framework for improving the estimate of Kalman gain, however, the high computation cost presents challenge for practical application. In this chapter, we introduce a bootstrap version of hierarchical filter that improves the robustness of the estimate of Kalman gain and also substantially reduces the computation cost of the hierarchical filter in its original version as the challenge of evaluating a large number of realizations is avoided by using

bootstrap.

4.1 Bootstrap concepts

Bootstrap is a nonparametric computer-intensive resampling method for statistical inference. Bootstrapping uses repeated samples from the parent data set to compute the statistics of interest, such as, confidence intervals, bias, and variance of an estimator θ. For our problem, the parent data set is the forecast ensemble, which is actually a sample from the underlying probability distribution.

The augmented state vector that includes model parameters, state variables and the corresponding simulated data is defined as

$$\psi_i^f = \begin{bmatrix} y_i^f \\ d_i^f \end{bmatrix}, \qquad i = 1, 2, \ldots, N_e \ .$$

Each augmented state vector contains $N_y + N_d$ entries (N_y is the dimension of y_i^f, and N_d is the dimension of d_i^f). The augmented forecast ensemble containing N_e augmented state vectors is denoted using Ψ^f. In the bootstrapping framework, we randomly sample from Ψ^f with replacement to generate N_B bootstrapped samples of the augmented forecast ensemble, Ψ^{f^*} that has the same ensemble size as the original ensemble Ψ^f. In this work, our interest lies in estimating the variances of θ, where θ is any quantity of interest. If the objective is to quantify the uncertainty in Kalman gain, θ denotes Kalman gain (K_e). If the goal is to assess the uncertainty associated with the covariance matrices, θ stands for C_{yd}^f or C_{dd}^f. The θ^* calculated from the N_B augmented forecast ensemble forms an empirical distribution, which is an estimate of the underlying unknown theoretical distribution of θ. For each element in θ, the plug-in estimate of variance is calculated as

$$\hat{\sigma}^2_{\theta_{i,j}} = \frac{\sum_{m=1}^{N_B} (\theta^*_{i,j,m} - \bar{\theta}_{i,j})^2}{N_B}, \tag{4.1}$$

and subsequently, the squared variation coefficient is defined as the ratio of the variance to the squared mean

$$\hat{C}^2_{v_{i,j}} = \frac{\hat{\sigma}^2_{\theta_{i,j}}}{\bar{\theta}^2_{i,j}} \ , \tag{4.2}$$

where the lower and upper limits of subscripts i and j depend on the definition of θ. If θ denotes C^f_{yd} or K_e, $i \in [1, N_y]$ and $j \in [1, N_d]$. If θ denotes C^f_{dd}, $i \in [1, N_d]$ and $j \in [1, N_d]$. In this chapter, we concentrate on eliminating the spurious correlations in the estimate of Kalman gain, the denoising process on covariances will be discussed in next chapter. Thus, θ denotes K_e in this chapter. About the value of N_B, there is no specific requirement. The larger the number of bootstrapped samples, the more reliable the estimate of variation coefficient will be, but $N_B = 50$ is often enough to give a good estimate of standard error (Efron and Tibshirani, 1993).

Bootstrapping allows one to gather many alternative versions of the single statistic. The empirical distribution of the bootstrap statistic and $\hat{C}^2_{v_{i,j}}$ are measures of the reliability of θ. Such information can be used to reduce the magnitude of the unreliable entries in θ, which can be achieved through the element-wise multiplication of θ with factor that is inferred from the empirical distribution of θ^*. Since the function of the factors is to screen out unreliable entries, the factors are termed as screening factor in this work. In the rest of the chapter, two more ways of defining screening factors are derived followed by the bootstrap version of hierarchical filter.

4.2 Bootstrapped version of hierarchical filter

Anderson (2007) proposed a hierarchical filter for reducing the effect of spurious correlations on the estimate of the Kalman gain, in which the confidence factors in the regression coefficients (similarly to components of the Kalman gain) are estimated from a group of independent ensembles of model realizations. In other applications (Vallès and Nævdal, 2008), the confidence factor has been applied directly to the component of the Kalman gain instead of the regression coefficient; so, for simplicity,

that is how it is applied here. In this method, m groups of N_e-member ensembles are generated and an estimate of the Kalman gain is computed for each ensemble. A confidence factor, for the entry (i, j) in the Kalman gain, is then defined as the value of $\alpha_{i,j}$ that minimizes the expression

$$\sum_{p=1}^{m} \sum_{q=1, q\neq p}^{m} (\alpha_{i,j} K^p_{e_{i,j}} - K^q_{e_{i,j}})^2.$$

The optimal value for $\alpha_{i,j}$ in the hierarchical filter is

$$\alpha_{i,j} = \frac{m - R^2_{i,j}}{(m-1)R^2_{i,j} + m} \ ,$$

where $R^2_{i,j} = \hat{\sigma}^2_{k_{i,j}} / \bar{K}^2_{e_{i,j}}$ is squared variation coefficient. The reason of using $R^2_{i,j}$ instead of $\hat{C}^2_{v_{i,j}}$ is that the way of calculating variance ($\hat{\sigma}^2_{K_{i,j}}$) is not plug-in estimate as shown in Eq. 4.1, because the mean $\bar{K}_{e_{i,j}}$ is unknown and an unbiased estimate of variance should be used, in other words, the denominator is $m - 1$ not m. Anderson (2007) suggests that $\alpha_{i,j}$ be truncated so that it does not take negative values. The purpose of the multiple ensembles is to provide estimates of mean and variance of K_e. The confidence factors provide an assessment of the accuracy of the correlations present in the Kalman gain matrix. Small value of $\alpha_{i,j}$ suggests that the correlation of the corresponding state variable with the data is unreliable and thus should be eliminated or reduced in magnitude. Hence, in this work, we would like to refer to confidence factors as screening factors.

The weakness of the hierarchical filter is the high computation cost of propagating the multiple N_e-member ensembles. In the proposed bootstrap version of the hierarchical filter method, we treat the original ensemble as the population and randomly resample with replacement to generate N_B bootstrapped ensembles. The objective behind the bootstrap resampling in the current study is to assess the accuracy of Kalman gain, as in the hierarchical filter, but without the cost of generating additional ensembles.

A bootstrapped sample of the Kalman gain matrix, K_e^*, is computed from each of the N_B resampled ensembles. By minimizing the same form of objective function as used in Anderson (2007), the screening factor for the entry (i, j) in the Kalman gain matrix is computed as,

$$\alpha_{i,j} = \frac{1 - \hat{C}^2_{v_{i,j}}/(N_B - 1)}{1 + \hat{C}^2_{v_{i,j}}} \ . \tag{4.3}$$

The detail derivation is provided in Appendix A.1. We use the expected bootstrap mean ($E[K_e^*] = K_e$) instead of the sample bootstrap mean ($\frac{1}{N_B}\sum_{p=1}^{N_B} K_e^{*^p}$) as the estimate of the population mean of the Kalman gain matrix. The bootstrapped samples of Kalman gain are, however, used to estimate the variance of the population using Eq. 4.1. The estimate of $\alpha_{i,j}$ will be positive for $N_B > \hat{C}^2_{v_{i,j}} + 1$. Although the possibility of negative values for $\alpha_{i,j}$ can be reduced through the use of a large N_B, we follow the suggestion of Anderson (2007) for the hierarchical filter and truncate the negative values to zero.

Once we have the computed screening factors of all elements, the screened Kalman gain is computed by multiplying the screening factors with the estimate of Kalman gain obtained from the standard EnKF in an element-wise manner:

$$K_e^s = \alpha \circ K_e \ ,$$

where $\circ$ denotes a Schur or Hadamard product. Following the screening of the original Kalman gain matrix, the standard updating (or analysis) step is carried out.

4.3 Alternative screening algorithms using bootstrap

One disadvantage of the previous method is that the optimal choice of $\alpha_{i,j}$ is sometimes negative, in which case the value is truncated at $\alpha_{i,j} = 0$. In order to avoid truncation, we put regularization on the estimate of screening factor. Thus, an alternative screening factor that is defined as the value of α that minimizes the following objective function

$$S(\alpha) = S_k(\alpha) + S_\alpha(\alpha) \ , \tag{4.4}$$

where $S_k(\alpha)$ is a measure of the difference between the screened estimate of Kalman gain and the true Kalman gain. Since we do not know the true Kalman gain, we use the expected bootstrap mean to approximate the true Kalman gain, then $S_k(\alpha)$ is expressed using Frobenius matrix norm as following

$$S_k(\alpha) = \frac{1}{2N_B} \sum_{p=1}^{N_B} \parallel (\alpha \circ K^*_{e_p} - K_e) \circ \lambda_k \parallel^2_F \ , \tag{4.5}$$

where λ_k is a matrix composed of the reciprocal of standard deviation for each entry $(1/\hat{\sigma}_{k_{i,j}})$.

$S_\alpha(\alpha)$ is a regularization term on the estimation of α that can be simple or with some complexity. If the variables to be updated are spatially correlated and preserving smoothness is desirable, $S_\alpha(\alpha)$ could be defined to minimize the magnitude of the derivative of α. In the following sections, we will introduce two types of $S_\alpha(\alpha)$.

4.3.1 Using a simple regularization term

$$S_\alpha(\alpha) = \frac{1}{2} \parallel \alpha \circ \lambda_\alpha \parallel^2_F \ , \tag{4.6}$$

where λ_α is a matrix containing $1/\sigma_\alpha$ for all entries. σ_α is a weighting factor for regularizing the estimation of α.

After substituting Eqs. 4.5 and 4.6 into Eq. 4.4, the 2nd derivative of $S(\alpha)$ is seen to be positive definite. Thus, the least square solution for screening factor is obtained by differentiating Eq. 4.4 with respect to α and equating it to zero. (Detailed derivation is included in Appendix A.2.) To avoid confusion, we use α_r to denote the regularized point-wise estimate of screening factor,

$$\alpha_{r_{i,j}} = \frac{1}{1 + (1 + 1/\sigma^2_\alpha)\hat{C}^2_{v_{i,j}}} \ . \tag{4.7}$$

When a regularization term is used in the definition of the screening factor, the solution depends on both the squared variation coefficient, $\hat{C}^2_{v_{i,j}}$, and on the value chosen for σ^2_α. Although many criteria could be used to select α, this particular objective function $S(\alpha)$ has the property that if the variance of the estimate of the Kalman gain is small, then the first term will be heavily weighted and the optimal α will be approximately 1. If the variance of the estimate of the Kalman gain is large compared to the numerator, then the second term is weighted more heavily and α will be approximately 0.

The effectiveness of eliminating spurious correlations increases as the value of σ^2_α is decreased, but, at the same time, the possibility of removing true correlations also increases. There is a tradeoff between the benefit of eliminating spurious correlations and the harm done by removing true correlations that must be balanced when selecting the value for σ^2_α. Cross-validation has been used in previous studies to select the optimal values of shrinkage parameters for covariance estimation (Friedman, 1989). It would be possible to use cross-validation in a similar way to select a value of σ_α that minimizes the RMSE in the estimate of model variables, but in this study we selected σ_α as an ad-hoc balance of reduction in spurious correlation against reduction of true correlations. Clearly, it is advantageous for α to be approximately equal to 0 when the correlation between data and model variables is small. The choice of σ_α will be addressed in the 1-dimensional linear example.

4.3.2 Using a smoothing regularization term

In the previously derived two expressions (Eq. 4.3 and Eq. 4.7), the screening factors are estimated individually without considering the smoothness of estimates. The estimated screening factors may fluctuate spatially. Instead of point-wise estimate, we can include the neighboring screening factors into the estimation process by minimizing the magnitude of the derivative of α. $S_k(\alpha)$ remains unchanged while $S_\alpha(\alpha)$ is

expressed as

$$S_\alpha(\alpha) = \frac{1}{2}(\alpha^T(W^TW + \frac{1}{\sigma_\alpha^2}I)\alpha) \ , \tag{4.8}$$

where I denotes identity matrix, W is an approximated second order derivative operator matrix (Oliver et al., 2008). Substituting Eq. 4.8 into Eq. 4.4. Taking the 1st order derivative of Eq. 4.4 and setting it equal to zero, we obtain

$$(W^TW + \Lambda)\alpha_s = \gamma \ , \tag{4.9}$$

where $\gamma = [\frac{1}{\hat{C}_{v_1}^2}, \frac{1}{\hat{C}_{v_2}^2}, \cdots, \frac{1}{\hat{C}_{v_{N_m}}^2}]^T$, the subscript N_m is the model dimension (for 2D or 3D model, N_m is the number of grids in xy plane), and α_s denotes the smooth estimate of screening factor, Λ is a diagonal matrix with each element $\Lambda_{i,i} = \frac{1}{\sigma_\alpha^2} + 1 + \frac{1}{\hat{C}_{v_i}^2}$.

A number of methods can be used to solve the linear system of equations given in Eq. 4.9. For a small model, Gaussian elimination could be a good choice, however, iterative methods can be a practical choice for a large model. The application on a 2D nonlinear problem showed that Gauss-Seidel iterative algorithm has better convergence than Jacobi algorithm for solving Eq. 4.9.

Fig. 4.1 illustrates the workflow of a general bootstrap-based screening algorithm. By bootstrapping the original ensemble, multiple replicates of Kalman gain are obtained and form an empirical distribution. Based on the empirical distribution, the squared variation coefficients ($\hat{C}_v^2$) are computed, which are finally used for calculating the screening factors. So far, we have introduced three different ways of defining the screening factors: point-wise estimate without regularization (Eq. 4.3), regularized point-wise estimate (Eq. 4.7), and estimate with smooth regularization (Eq. 4.9). These three types of screening factors all use the information from $\hat{C}_v^2$, but apply different level of regularization on the estimation of α, in other words, different level of prior information about α is introduced into the optimization process. Regardless of the type of regularization, the resulting screening factors are multiplied with the estimate of Kalman gain obtained from the standard EnKF in an element-by-element

manner. Improved estimate of Kalman gain (K_e^s) is obtained through screening as shown in the subfigure at the end of the flowchart.

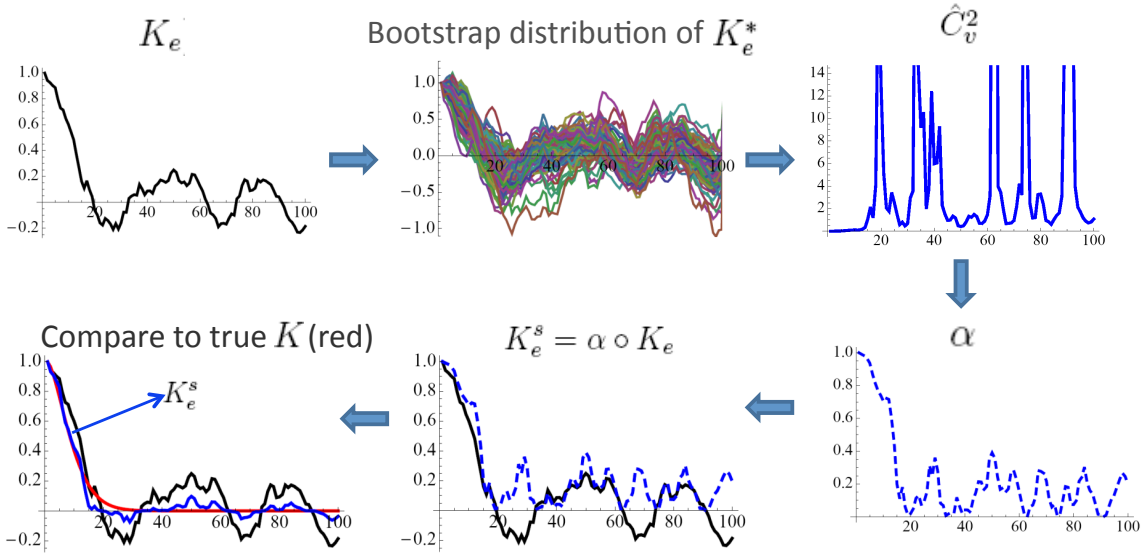


Figure 4.1: An illustration of the workflow of bootstrap-based screening algorithm.

4.4 Linear example

Consider a correlated 1D Gaussian random vector $Z = \{Z_1, \ldots, Z_{100}\}$ whose prior mean is 0 and whose covariance is only a function of the distance between entries, i.e. $\text{cov}(Z_i, Z_j) = C(|i - j|)$. The covariance function is in the exponential family with an exponent of 1.5 and a practical range of 40:

$$C(h) = \exp[-3.(|h|/40.)^{1.5}].$$

A single measurement, $d_{\text{obs}} = 2$, with additive Gaussian noise (standard error of 0.05) is made of Z_1. The initial ensemble is generated as independent, unconditional realizations from the same distribution as the prior for Z. For all the bootstrap based screening algorithms, the number of bootstrap replicates, N_B=100. Because this problem is relatively small, the true Kalman gain for this problem can be easily computed.

When the true covariance is known, the optimal value of screening factors can be computed using the method of Furrer and Bengtsson (2007, Eq. 20). Although the assumption of known covariance is generally impractical, it allows a useful evaluation of the effect of σ_α on the localization. Fig. 4.2 shows a comparison of the optimal localization for an ensemble size of 400 compared to the estimates of screening factors computed using Eq. 4.7 with σ_α ranging from 0.1 to 5. While small values of σ_α are effective at eliminating spurious correlations, they eliminate too much of the Kalman gain in the intermediate region. Based on the overall performance in regions of weak and strong correlations, we chose to use $\sigma_\alpha = 0.6$ for all screening with regularization (Eq. 4.7 and Eq. 4.9).

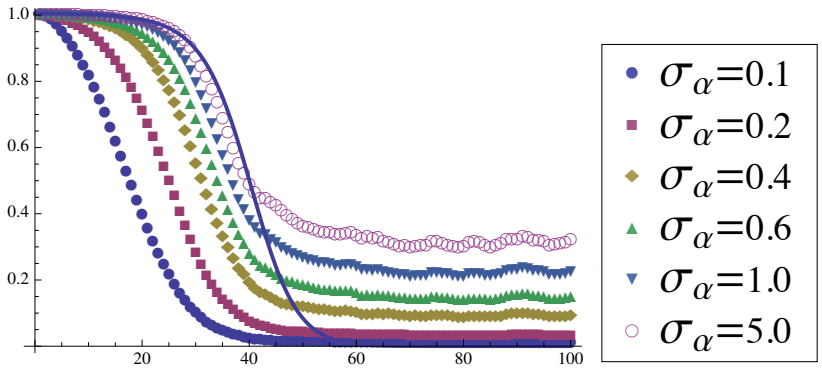
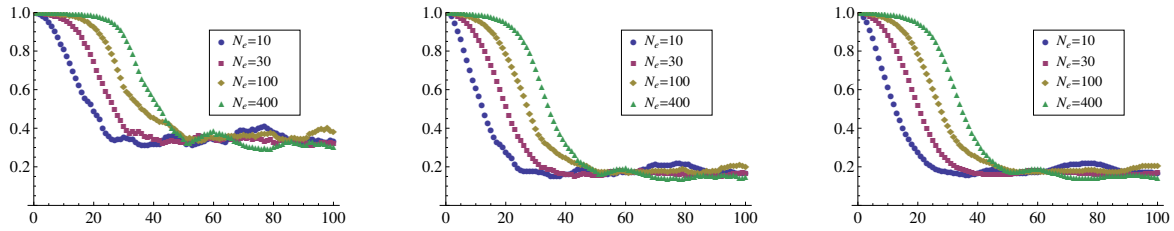


Figure 4.2: Expected values of the screening coefficient α for updating variables to an observation of the first variable with ensemble size of 400. The solid blue curve is the optimal localization of Furrer and Bengtsson (2007).

Fig. 4.3 shows the mean values of screening factors, computed from averages of 100 ensembles for four different ensemble sizes for the three methods that have been proposed here. It is observed that as the ensemble size increases, the size of the region of localization for this example also increases for all the three methods. (Not all problems would have spatially correlated variables.) Note, however, that the widths of the localization region are slightly wider for the point-wise estimate without regularization (Eq. 4.3) than the regularized point-wise estimate (Eq. 4.7) and the estimate with smooth regularization (Eq. 4.9). The other evident difference between

the estimate of screening factor shown in Fig. 4.3 (a) and those regularized estimates shown in Figs. 4.3 (b) and (c) is in the mean value of screening factor at large distances from the measurement location. In none of the methods does the mean value of screening factor approach zero. The value is approximately 0.15 for the regularized point-wise estimate (α_r) and the smooth estimate (α_s), while for point-wise estimate (α), the value is approximately 0.30. In this regard, the use of regularized estimates of screening factor would be preferred. The only difference between the regularized point-wise estimate (α_r) and the smooth estimate (α_s) is the level of smoothness.

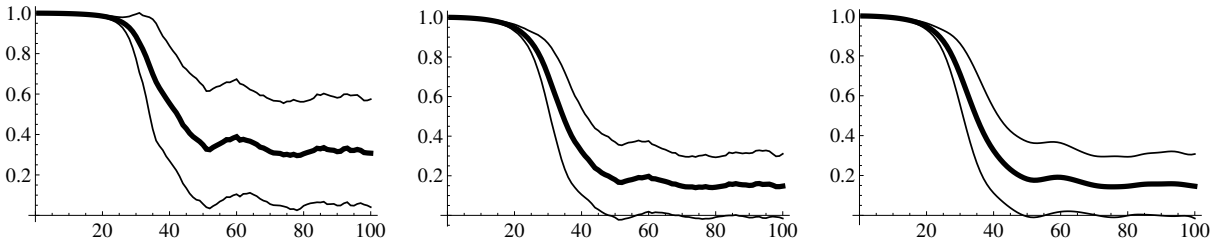


(a) Point-wise estimate (α) (b) Regularized point-wise estimate (α_r) (c) Smooth estimate (α_s)

Figure 4.3: Mean values of the screening factors computed for different ensemble size (N_e).

The benefit of using bootstrap method is that screening factor can be computed from a single ensemble, without the need to generate and evaluate multiple ensembles. When screening factor is computed from a single ensemble, however, the estimate may not be very close to the mean values shown in Fig. 4.3. Based on the screening factors calculated from each of the 100 ensembles, the variability of the estimates of screening factor was computed for the three methods. Fig. 4.4 compares the variability of the estimates of screening factor from the three methods. Note that the variability in screening factor is always small near the measurement location, although that would not be true for other types of data for which the true covariance is not distance-dependent.

The objective of estimating screening factor is to reduce the effect of spurious correlations on the Kalman gain. The correct Kalman gain is shown as a red curve



(a) Point-wise estimate with one standard deviation

(b) Regularized point-wise estimate with one standard deviation

(c) Smooth estimate with one standard deviation

Figure 4.4: Comparison of the variability shown in the three types of estimates of the screening factors for the case of $N_e = 400$.

in each of the subplots of Fig. 4.5. Note that the Kalman gain estimates from the ensemble are generally quite good in the vicinity of the measurement location, but that spurious correlations are present even when the Kalman gain should be zero. When the ensemble size is small (for the case of $N_e = 30$), significant spurious correlations are observed in the standard estimate of Kalman gain (black curve). Through screening, the resulting estimates of the Kalman gain are improved as the use of screening factors has removed most of the spurious correlations, but has slightly underestimated part of the true Kalman gain. In terms of the performance of eliminating spurious correlations, the EnKF with screened Kalman gain (EnKF-SKe) using α_r and EnKF-SKe (α_s) are better than EnKF-SKe (α).

Fig. 4.6 compares the mean Kalman gains with their associated variability for the standard EnKF and for the three EnKF-SKe methods, which are obtained based on 100 evaluations with different randomly generated initial ensemble. The standard estimate of the Kalman gain has the largest variability, but is unbiased (Fig. 4.6 (a)). The three screened estimates (Figs. 4.6 (b), (c) and (d)) all have smaller variability compared to the unscreened Kalman gain, but the average estimates are biased. Since here we use a very small ensemble size of 30, the standard estimate of Kalman gain for that intermediate region (where showing strong bias) often has smaller or larger magnitude than the correct Kalman gain, and may even has an opposite sign, which

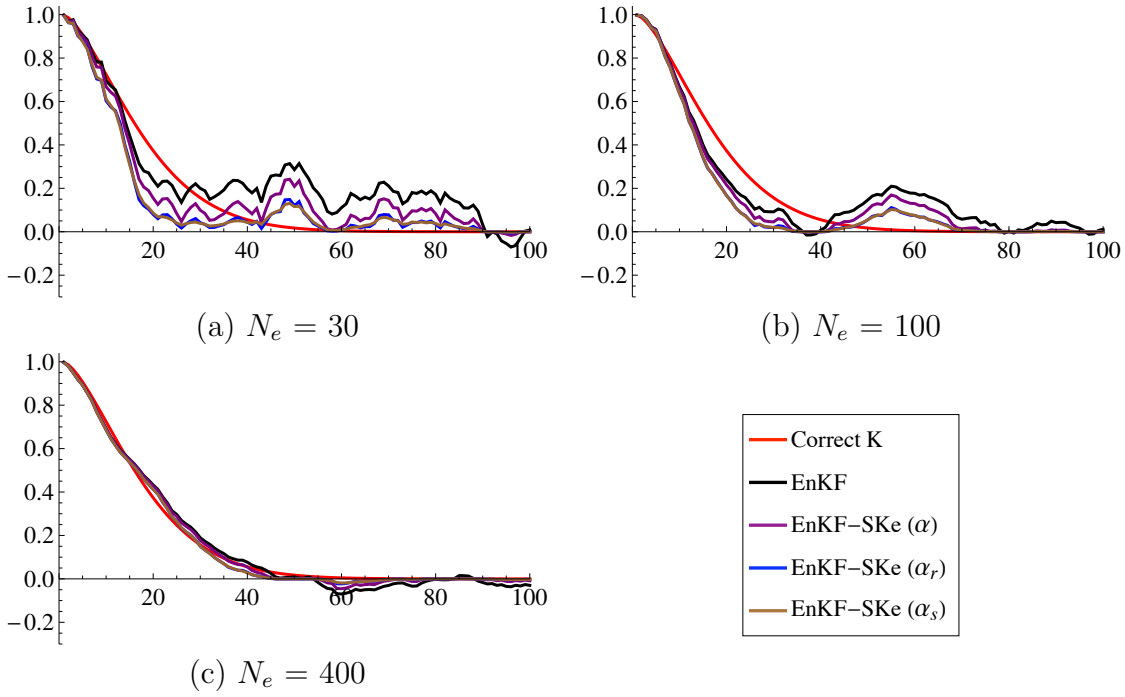
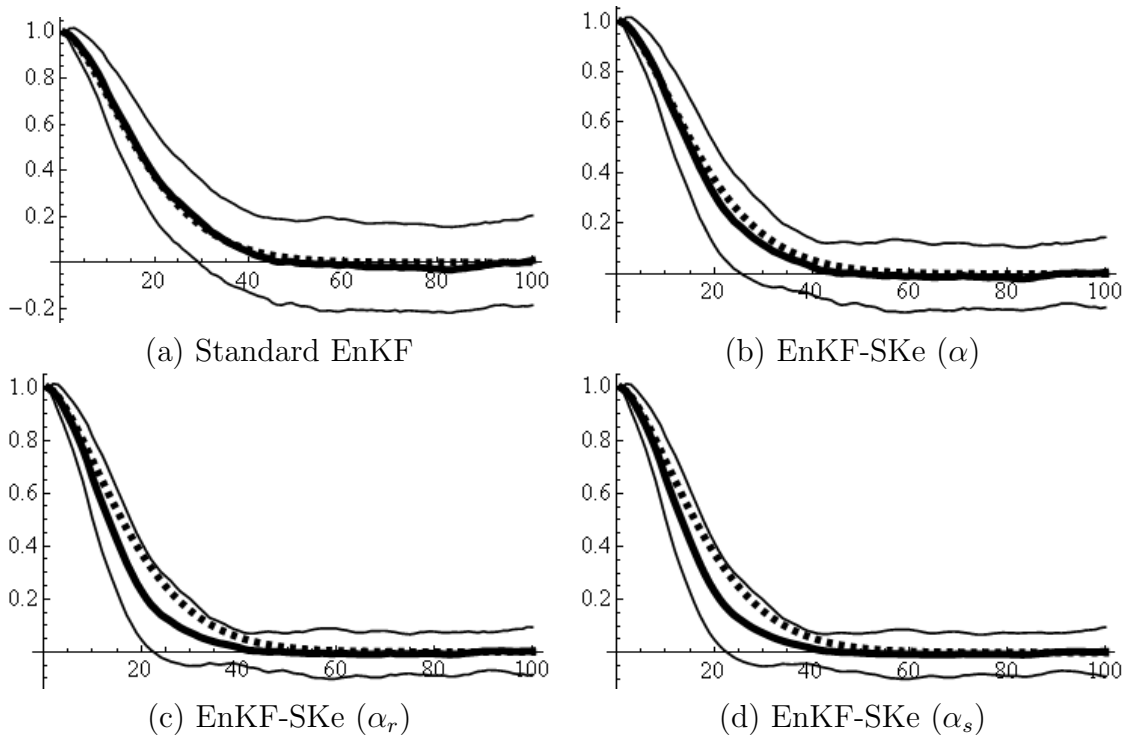


Figure 4.5: The estimates of Kalman gain obtained from different methods for the cases with different ensemble sizes (N_e).

indicates that the standard estimate for the region of low-level true correlations is unreliable and screening factor is most likely to be lower than 1.0. Multiplying the screening factors with values lower than 1.0 leads to the bias shown in the statistical mean of screened estimates from 100 ensembles. For the standard estimate that has different sign from the correct Kalman gain, or that has the same sign but larger magnitude than the correct Kalman gain, the screened estimate gets closer to the correct Kalman gain through multiplying with the screening factors. In the case of negative biased standard estimate (that has the same sign but smaller magnitude than the correct Kalman gain), multiplying with screening factors results in an estimate that is even further from the correct Kalman gain. In the uncorrelated region, the standard errors vary from 0.2 (standard EnKF), to 0.15 (EnKF-SKe (α)), to 0.08 (EnKF-SKe (α_r) and EnKF-SKe (α_s)).

We further investigated the performance of the EnKF-SKe methods on a history



(a) Standard EnKF (b) EnKF-SKe (α)

(c) EnKF-SKe (α_r) (d) EnKF-SKe (α_s)

Figure 4.6: Mean estimate of the Kalman gain with one standard deviation for the case of $N_e = 30$. Dashed curve in all subfigures shows the correct Kalman gain.

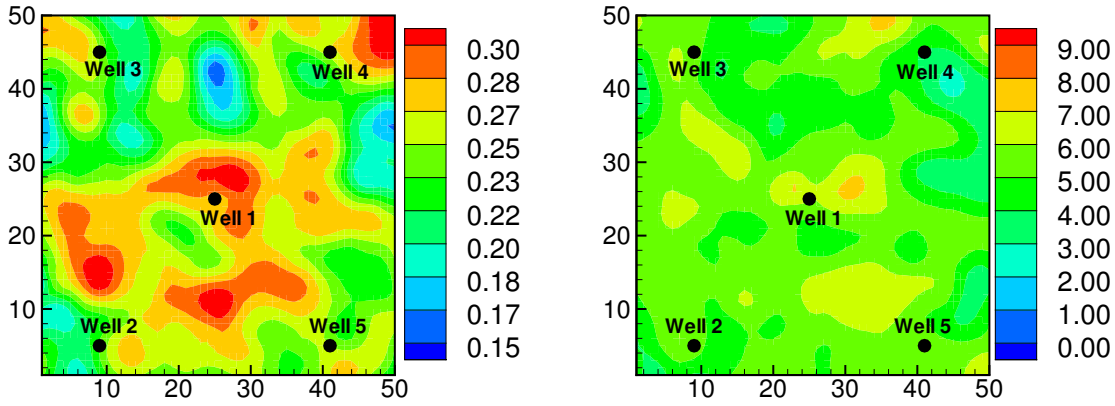
matching problem of a 2-dimensional, 2-phase reservoir model. The results from the proposed methods are also compared with the results obtained from the implementation of the standard EnKF on the same history matching problem.

4.5 Comparison study on the 2-dimensional, 2-phase reservoir model

The EnKF-SKe methods are applied to a 2-dimensional water flooding reservoir history matching problem. In order to analyze the ability of the newly proposed method for screening out the spurious correlations, the standard EnKF is also used on the same history matching problem and the performances from the EnKF with/without screening are compared. In case of a multiphase reservoir history matching problem, the relationship between the production data and the model parameters (e.g. porosity and permeability) is highly nonlinear during transient flow period.

4.5.1 Model description and production profiles from the reference model

The synthetic "true" reservoir model has grid dimensions of $50 \times 50 \times 1$ and an individual gridblock dimension is 30 ft $\times$ 30 ft $\times$ 20 ft. There are 4 producers and 1 injector in the field. Both the log permeability and the porosity fields are generated using isotropic Gaussian variogram models with a practical range of 8.5 ($\approx 15/\sqrt{3}$) grids. The resulting porosity and log permeability fields are multivariate Gaussian with means of 0.25 and 5.2 and standard deviations of 0.03 and 0.8, respectively. The coefficient of correlation between the log permeability and porosity is 0.5. Fig. 4.7 shows the contour plots of the reference log permeability and porosity fields. In these figures, the black points indicate the locations of the five wells in the field. Table 4.1 summarizes the exact well locations, the primary constraints, and the secondary production constraints on individual well.



(a) Porosity (b) Log permeability

Figure 4.7: The true porosity and log permeability fields.

The reference reservoir model is produced for a total of 520 days. Fig. 4.8 shows the profiles of the production data from the reference model which include the oil and water production rates for the four producers as well as the bottom hole pressure for the injector.

Table 4.1: Description of the wells ("-" denotes the same specifications are used for the rest of the wells as those are used for well 2).

Well	1	2	3	4	5
x location	25	9	9	41	41
y location	25	5	45	45	5
Well type	injector	producer	-	-	-
Constraints	4000 stb/day	1000 psia	-	-	-
Limits	6000 psia	4500 stb/day	-	-	-

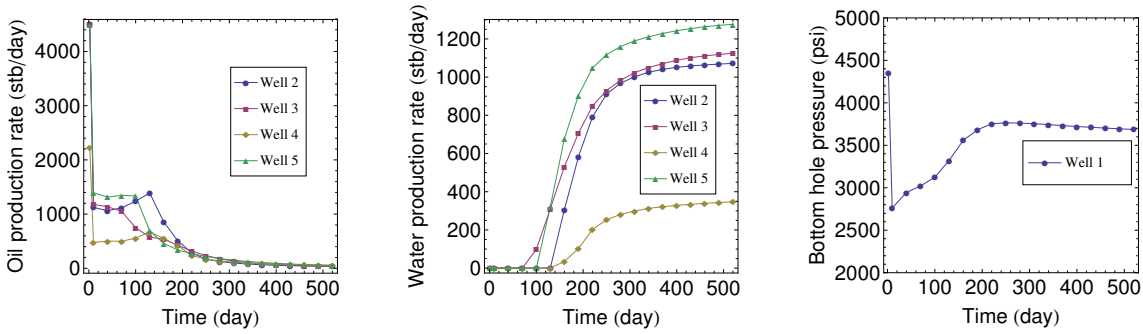


Figure 4.8: The production profiles of the true model.

4.5.2 Data assimilation setup

The total production period for this reservoir model is 520 days out of which the period of the first 160 days is treated as the production history and the period from day 161 to day 520 is treated as the prediction period. The water injection was started from day 0 and was continued until the end of the production period (520 days). The oil and water production rate data from each producer and the bottom hole pressure data of the injector are used as the observations during data assimilation. These observations are taken at day 2, day 10, and thereafter every 30 days until day 160. Thus, there are a total of 7 data assimilation time steps and 9 production data at each assimilation step. The measurement noise for the injector bottom hole pressure is assumed to have a mean and a standard deviation of 0 psi and 3 psi, respectively. The measurement noise for the oil production rate has a mean of 0 stb/day and a standard deviation of 3 stb/day. For water production rate data, the standard

deviation of measurement error is assumed to be 1% of the actual observation value.

In the current study, we wanted to verify the ability of the proposed EnKF-SKe methods for removing the spurious correlations. Therefore, a small ensemble of size 30 was used for introducing significant sampling errors. We included porosity, log permeability, pressure, and water saturation data for each gridblock into the state vector. Thus, each state vector contained 10,000 state variables.

4.5.3 Results and discussions

4.5.3.1 History matching production data

The ensemble of 30 realizations was continually updated by assimilating 9 production data at each data assimilation time step. We had 7 data assimilation time steps and thus, a total of 63 observations were assimilated during the entire history matching process. Once the data assimilation was complete over all the data assimilation time steps, the final updated ensemble of porosity and log permeability was evaluated from the beginning (day 0) up to the end of production period (day 520) using a commercial reservoir simulator. Fig. 4.9 through Fig. 4.11 show the results of the production data obtained by rerunning the final updated ensembles that are obtained from the standard EnKF and the three EnKF-SKe methods with $N_B = 50$. For reference, we also run a case of standard EnKF with $N_e = 1000$. Since the effect of spurious correlations should be small in an ensemble of this size, it provides a good basis for comparison.

In Fig. 4.9 through Fig. 4.11, the red dots denote the observations which were assimilated for estimating the model variables and the green dots denote the observations from the reference model during the prediction period. The observations from the prediction period were used for comparing the forecast results from the final ensembles resulted from different methods. The boxplot at each data assimilation time step summarizes the ensemble outputs. It is evident that the proposed EnKF-SKe methods provided better history matching results compared to the standard EnKF

with $N_e = 30$, especially for the water production rate data. Compare to the results of EnKF with $N_e = 1000$, EnKF-SKe methods result in a little biased predictions for some production data.

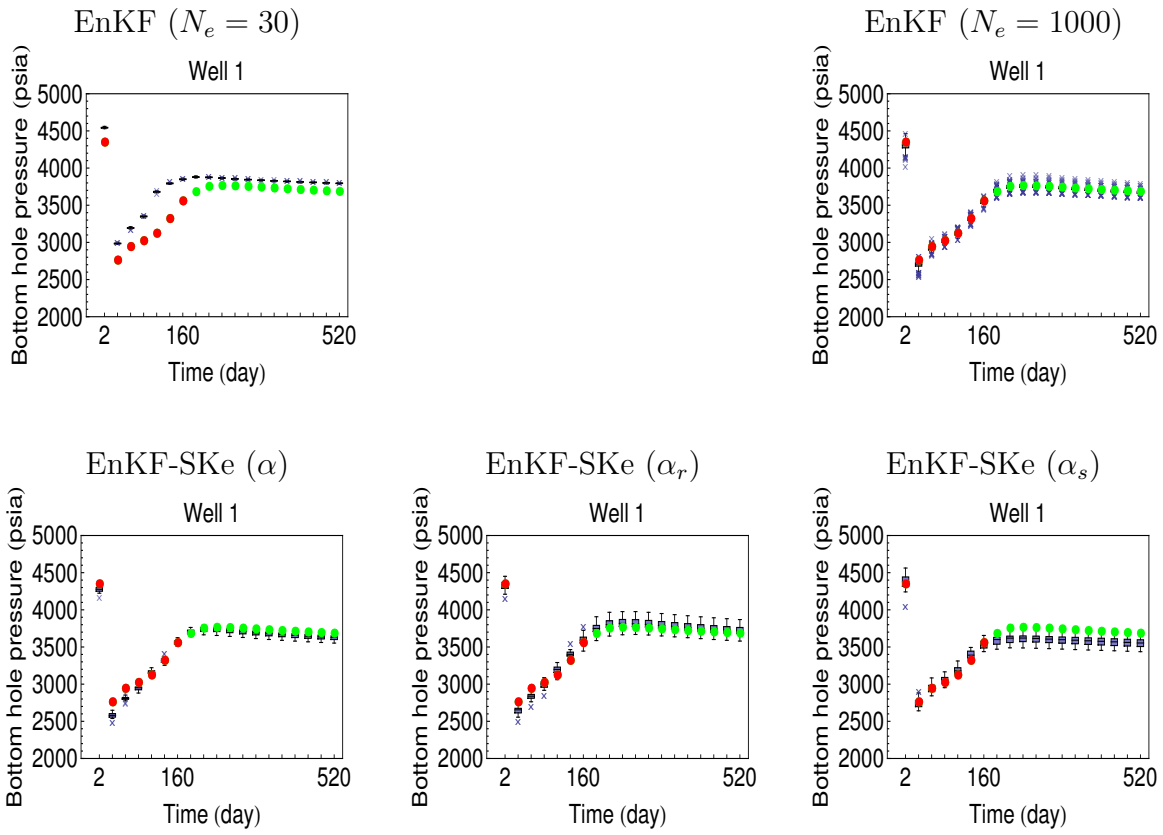


Figure 4.9: Comparison of production profiles: bottom hole pressure of the injector. (The observations are denoted by red dots (used for assimilation) and green dots (for reference in prediction period).)

4.5.3.2 Investigation of the Kalman gain

The only difference between the standard EnKF and the EnKF-SKe methods is the manner in which the Kalman gain is computed. In EnKF-SKe methods, the Kalman gain is screened using the screening factors whereas in the standard EnKF, there is no postprocessing of the Kalman gain. In order to explore the features of Kalman gain matrices obtained from these methods, we carried out additional investigations on the Kalman gain.

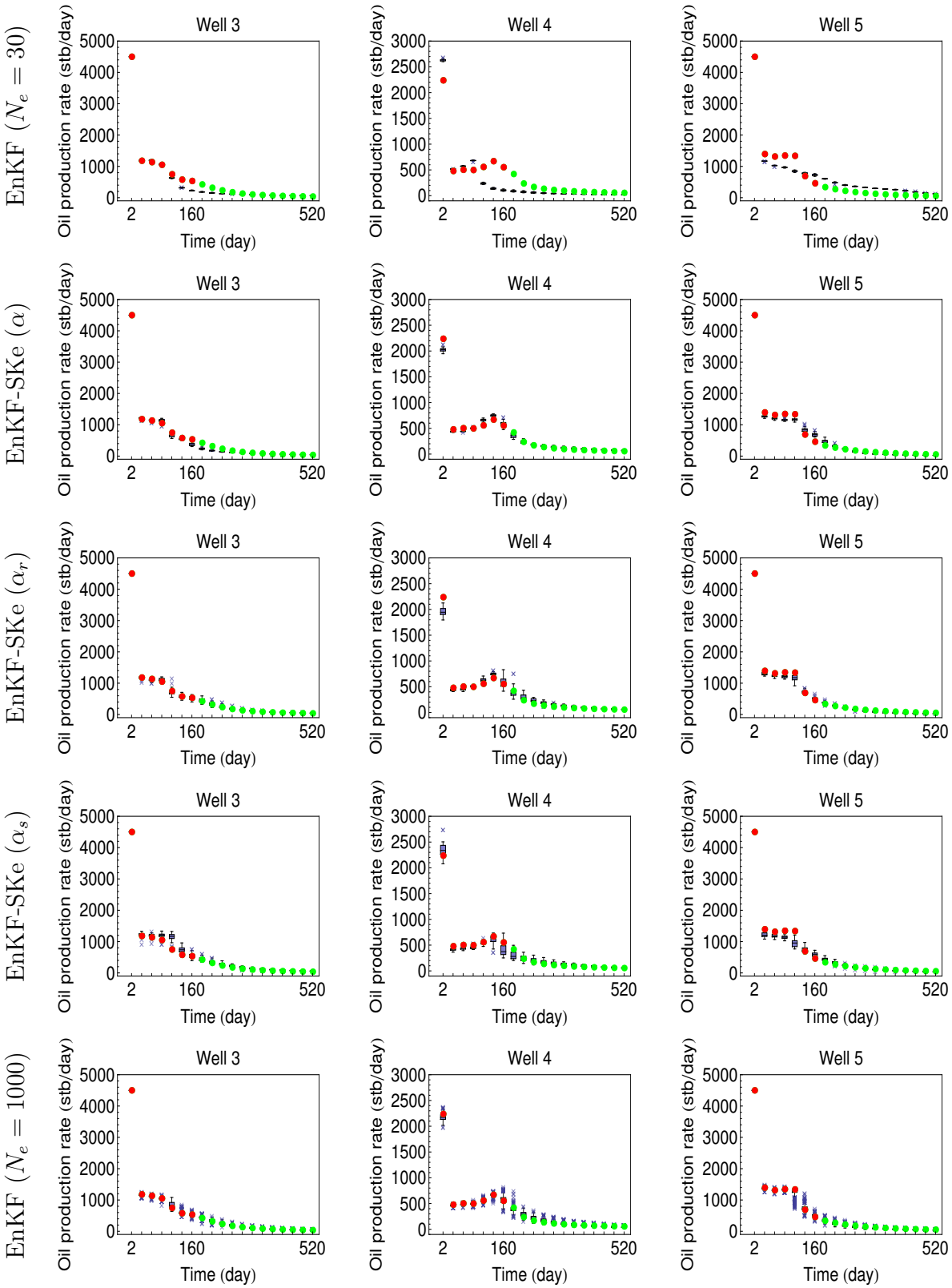


Figure 4.10: Comparison of production profiles: oil production rate. (The observations are denoted by red dots (used for assimilation) and green dots (for reference in prediction period).)

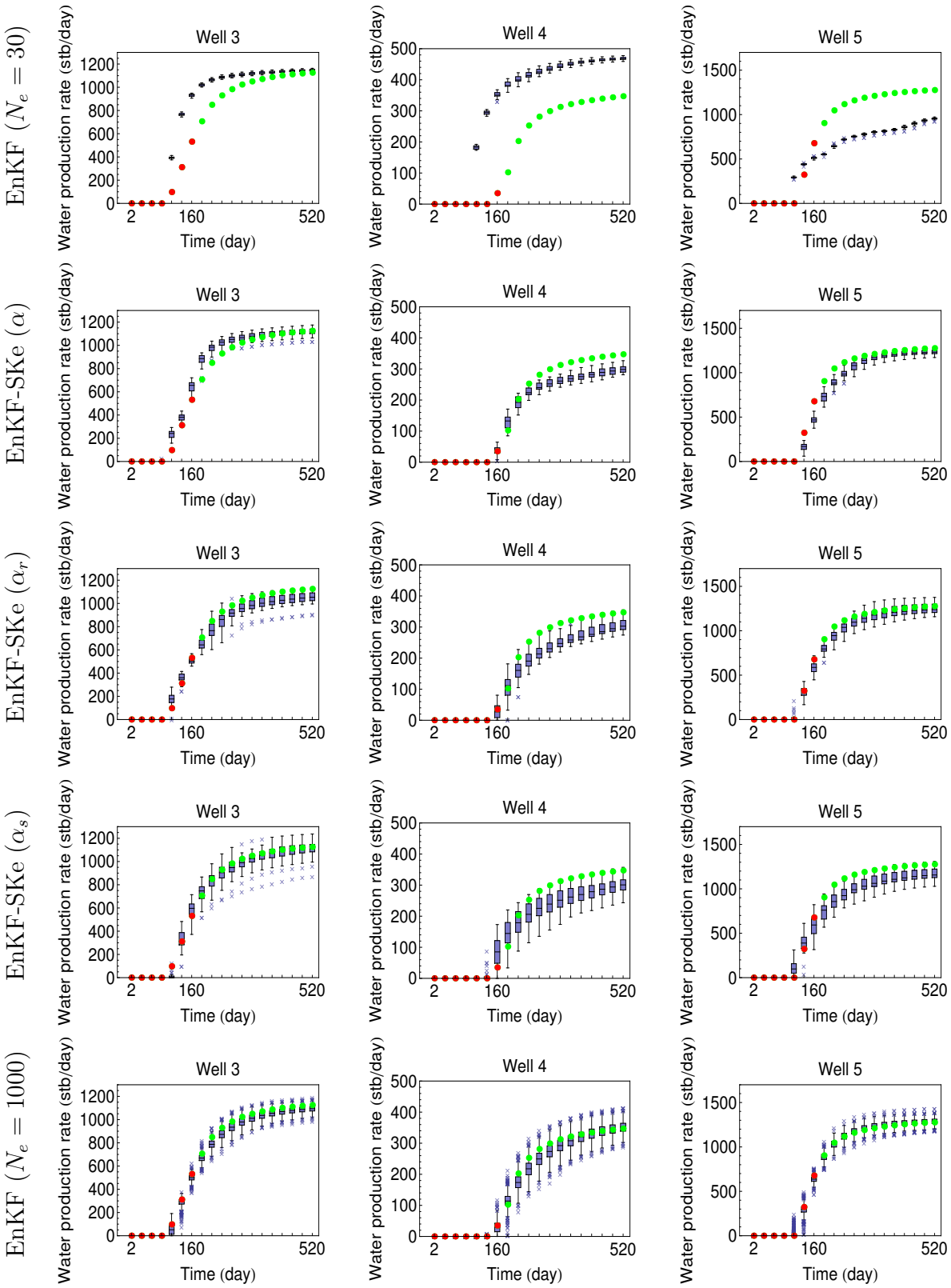
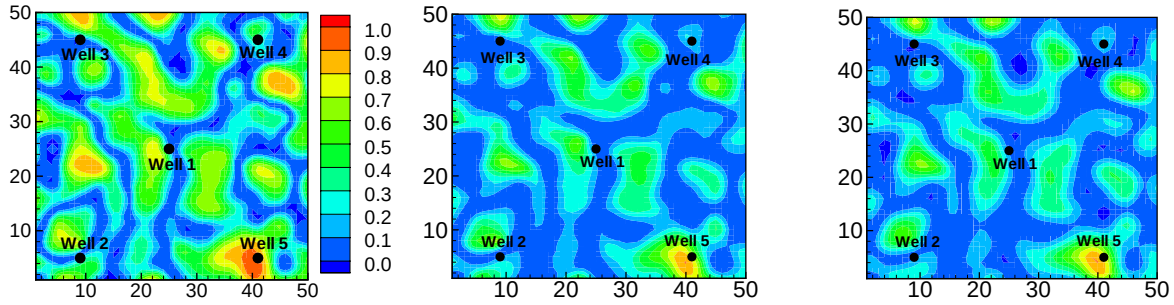


Figure 4.11: Comparison of production profiles: water production rate. (The observations are denoted by red dots (used for assimilation) and green dots (for reference in prediction period).)

Each element of the Kalman gain matrix is proportional to the correlation between state variables and predicted data, but is inversely proportional to the uncertainty of forecasted data and observation noises. Each column from the Kalman gain matrix corresponds to one datum or a measurement and each element in a column corresponds to a state variable. Every element from the Kalman gain matrix can be thought of as a weight that gets multiplied with the data mismatch, finally resulting into the increments (updates) to the corresponding state variable. For the present history matching problem, the 4th column from the Kalman gain matrix corresponds to the oil production rate (OPR) measurement at well 5. Fig. 4.13 and Fig. 4.14 show contour maps of the Kalman gain matrix which correspond to the OPR data of well 5 and different state variables at two different data assimilation time steps (1st and 7th). The significant difference observed in Fig. 4.13 between the Kalman gains obtained from the EnKF and EnKF-SKe methods is that the Kalman gain obtained from the EnKF-SKe methods showed greater numbers of regions with a value of zero as compared to the standard EnKF. The estimates of Kalman gain obtained from EnKF-SKe (α_r) and EnKF-SKe (α_s) show a smaller amount of spurious correlations than that from EnKF-SKe (α). Fig. 4.12 shows the screening factors multiplied with the Kalman Gain that is corresponding to the OPR data of well 5 and log permeability. The values shown in the point-wise estimate (α) are much higher than the values shown in the other two estimates.

It is evident that the EnKF-SKe methods are successful in removing most of the unrealistic correlations. However, the methods are clearly not able to eliminate them completely, as can be seen by comparing to the maps from EnKF with $N_e = 1000$. Fig. 4.14 shows the estimates of Kalman gain at the 7th data assimilation timestep. We can see the cumulative effects of removing spurious correlations at the 6 previous data assimilation timesteps. The effect is most evident in the column of pressure. The low-value region in the correlation map of EnKF ($N_e = 1000$) becomes a high-value



(a) Point-wise estimate (α) (b) Regularized point-wise estimate (α_r) (c) Smooth estimate (α_s)

Figure 4.12: The three types of estimates of screening factor multiplied with the Kalman Gain corresponding to the OPR data of well 5 and log permeability.

region of EnKF ($N_e = 30$), while we do not observe such large difference in the three maps of EnKF-SKe.

The results for the other columns from the Kalman gain matrix which correspond to other different types of data were also analyzed. The results showed the same characteristics as discussed in the case of Fig. 4.13 and Fig. 4.14 and therefore, they are not included here.

4.5.3.3 Model parameter estimates

Fig. 4.15 shows the estimates of log permeability fields obtained from the standard EnKF and the EnKF-SKe methods. Compared to the true log permeability field, the final estimate of the log permeability field provided by the standard EnKF with $N_e = 30$ does not look plausible as it contains a number of regions having extremely high values and regions with extremely low values, and these values are beyond the bounds of true log permeability field. Therefore, it is clear that the problem of overshooting occurred in the case of the standard EnKF with $N_e = 30$. The final estimates of the log permeability field from the EnKF-SKe methods are better than the standard EnKF with $N_e = 30$ as the final estimated log permeability values lie in the true range with only few locations having values beyond the limit values.

The grid-based standard deviations of the final updated model parameters (porosity and log permeability) among the ensemble members were calculated for all methods and shown in Fig. 4.16. The standard deviation values for the model parameters obtained from the EnKF-SKe methods were nearly 10 times greater than those obtained from the standard EnKF with $N_e = 30$. It was also observed that, the EnKF-SKe methods did not consume a great amount of variability as compared to the values of the initial standard deviation, which means that the EnKF-SKe method can substantially increase the effective rank of the ensemble. The scale shown in the STD maps of EnKF-SKe methods are comparable to those shown in the map from EnKF with $N_e = 1000$, but EnKF with $N_e = 1000$ shows a very clear pattern which cannot be found from the cases using a small ensemble size of 30, since the covariance does not only contain spurious correlations but also underestimated true correlations.

The root mean squared error (RMSE) of the final updated model parameters from the EnKF with/without screening were also computed and compared in Fig. 4.17. The RMSE values shown in the plot of EnKF with $N_e = 1000$ are generally lower than 2.0, while in the plots of EnKF-SKe methods the RMSE values are less than 2.0 for most of the regions except few regions with values around 4.0. The EnKF-SKe methods, however, resulted in a much better estimate of the model parameters than EnKF with $N_e = 30$, as the RMSE values from the EnKF-SKe methods were lower than those obtained from the standard EnKF with $N_e = 30$. We observed similar results for the final estimates of porosity field.

Fig. 4.18 shows the same three realizations of the final updated log permeability fields obtained from all the methods. Updated realizations from the large ensemble appear to be similar in character to the initial realizations and have maintained variability. The realizations obtained from the standard EnKF with $N_e = 30$ are all nearly identical because of the total loss of variability in the ensemble. The realizations obtained from the EnKF-SKe methods share some large-scale features,

but have maintained variablility. A close examination of these realizations shows a loss of smoothness, especially for EnKF-SKe (α_r). It is difficult to quantify the differences in the structure without analyzing the variograms (or covariances) for the resulting realizations.

Fig. 4.19 compares experimental variograms for the realizations from the large final ensemble ($N_e = 1000$) to realizations from the two smaller final ensembles ($N_e = 30$). Two differences are apparent. First, the sill is extremely high (4.6) for the realization from the standard EnKF with $N_e = 30$, compared to the sill (0.85) for the standard EnKF with $N_e = 1000$, or to the sill (1.35) from the EnKF-SKe (α_r) method with $N_e = 30$. Second, while the range of the variogram is less with EnKF-SKe (6.4) compared to either of the standard EnKF methods without screening (7.7 for $N_e = 30$ and 8.5 for $N_e = 1000$), the difference that is most apparent in the realizations appears to be a result of the change in character of the variogram near the origin. Instead of a Gaussian variogram that was used to generate the initial ensemble, the realization from EnKF-SKe has a variogram that is intermediate between Gaussian and exponential.

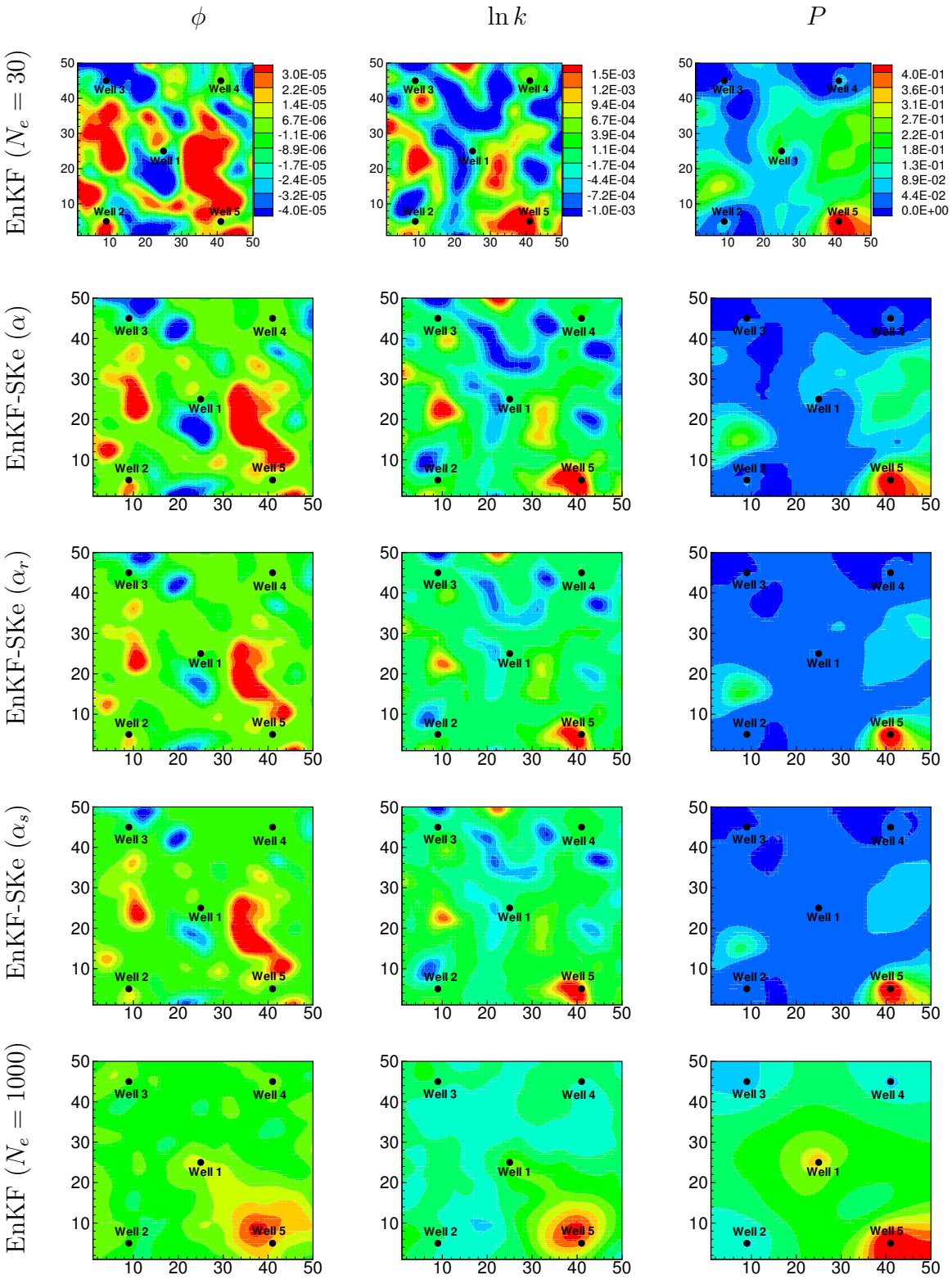


Figure 4.13: At the 1st data assimilation time step, the Kalman Gain matrix corresponding to the OPR data of well 5 and different state variables: ϕ (porosity), $\ln k$ (log permeability), P (pressure).

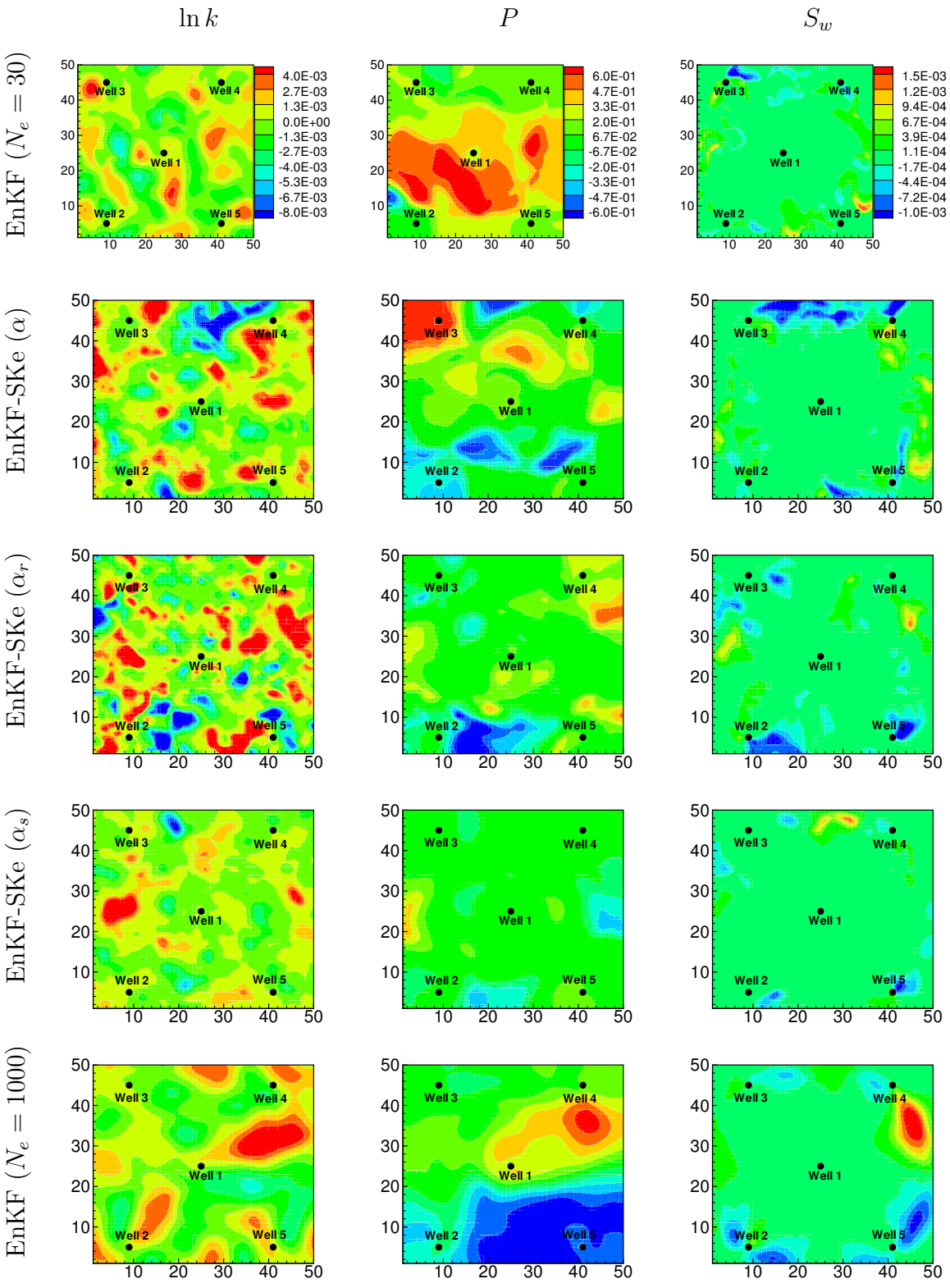


Figure 4.14: At the 7th data assimilation time step, the Kalman Gain matrix corresponding to the OPR data of well 5 and different state variables: $\ln k$ (log permeability), P (pressure), S_w (water saturation)

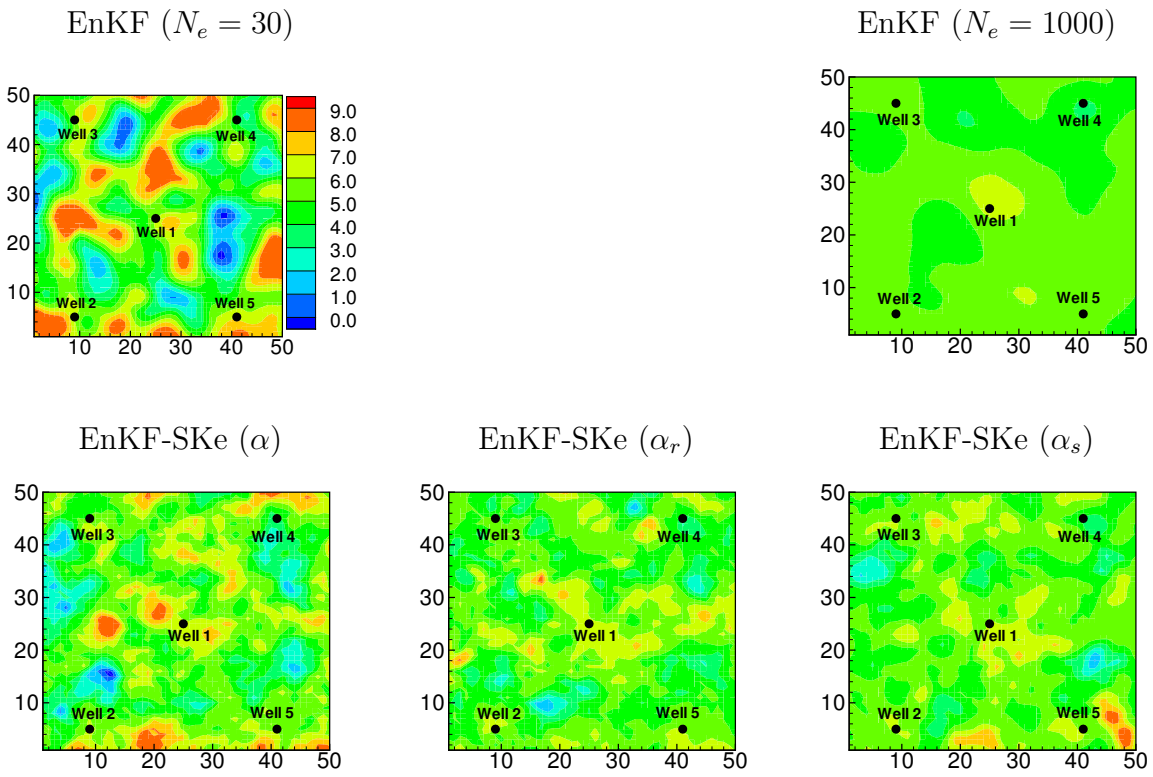


Figure 4.15: Final mean log permeability field.

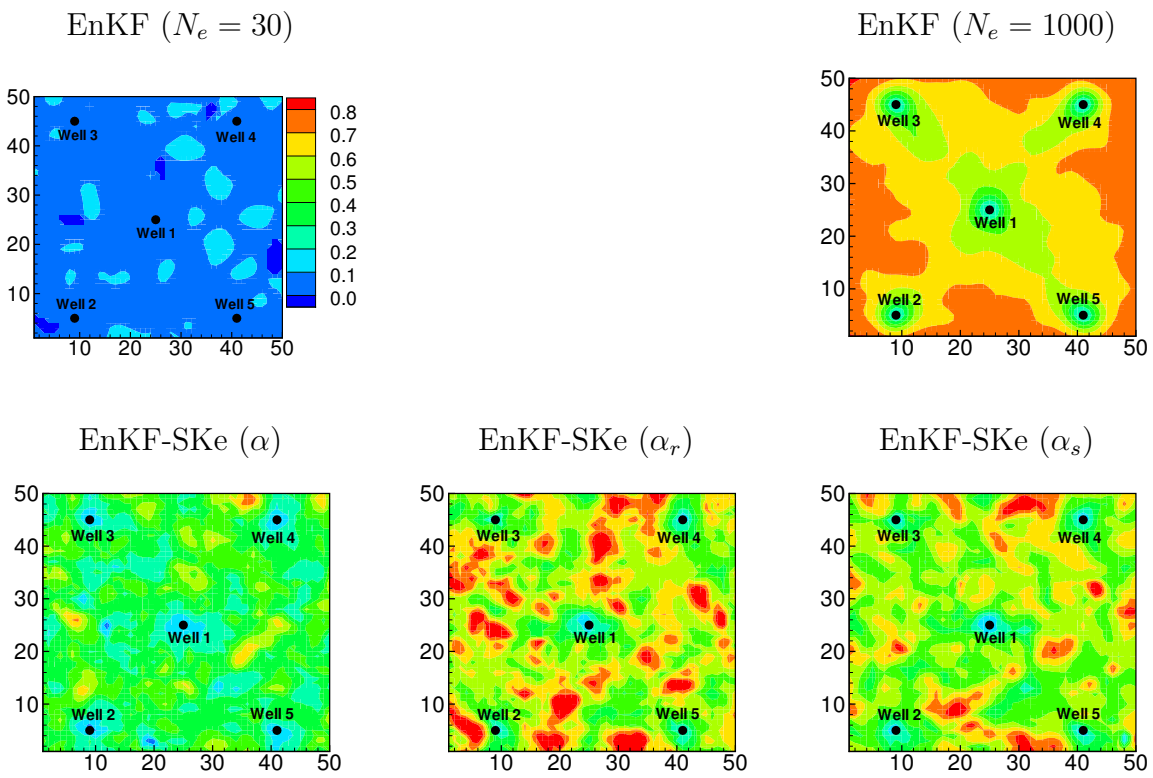


Figure 4.16: Final standard deviation of log permeability field.

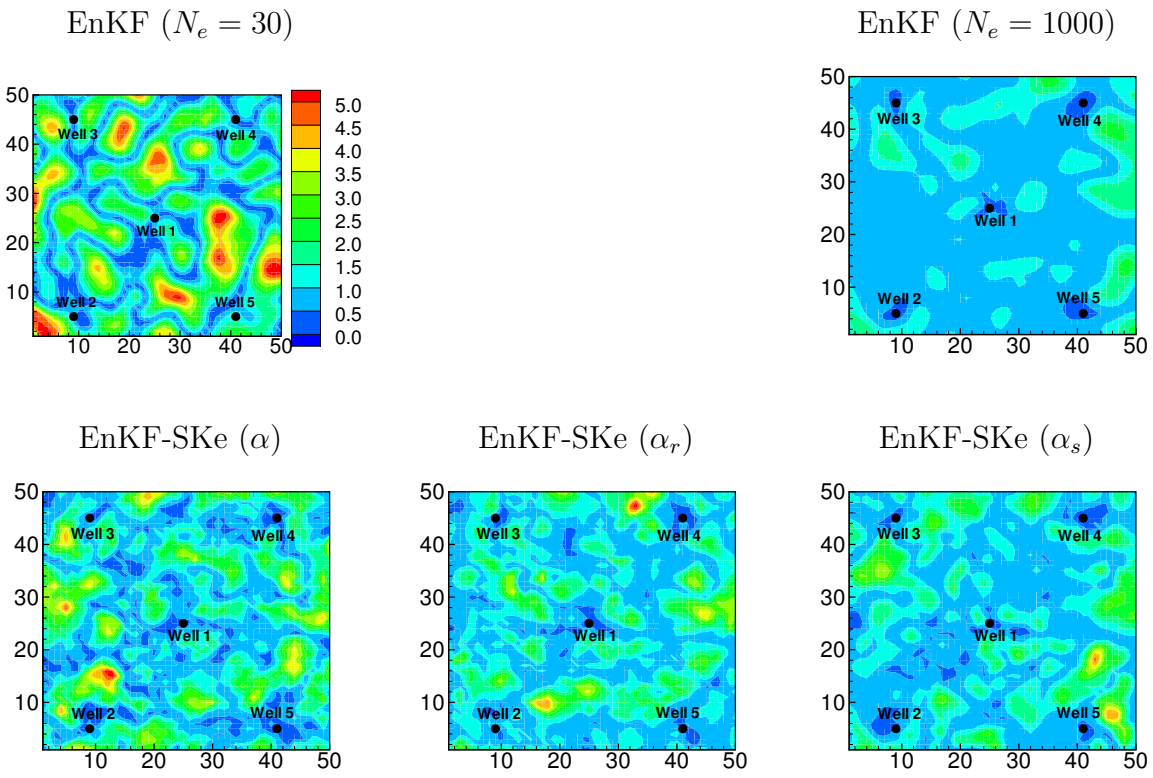


Figure 4.17: The root mean squared error (RMSE) of the estimate of log permeability field.

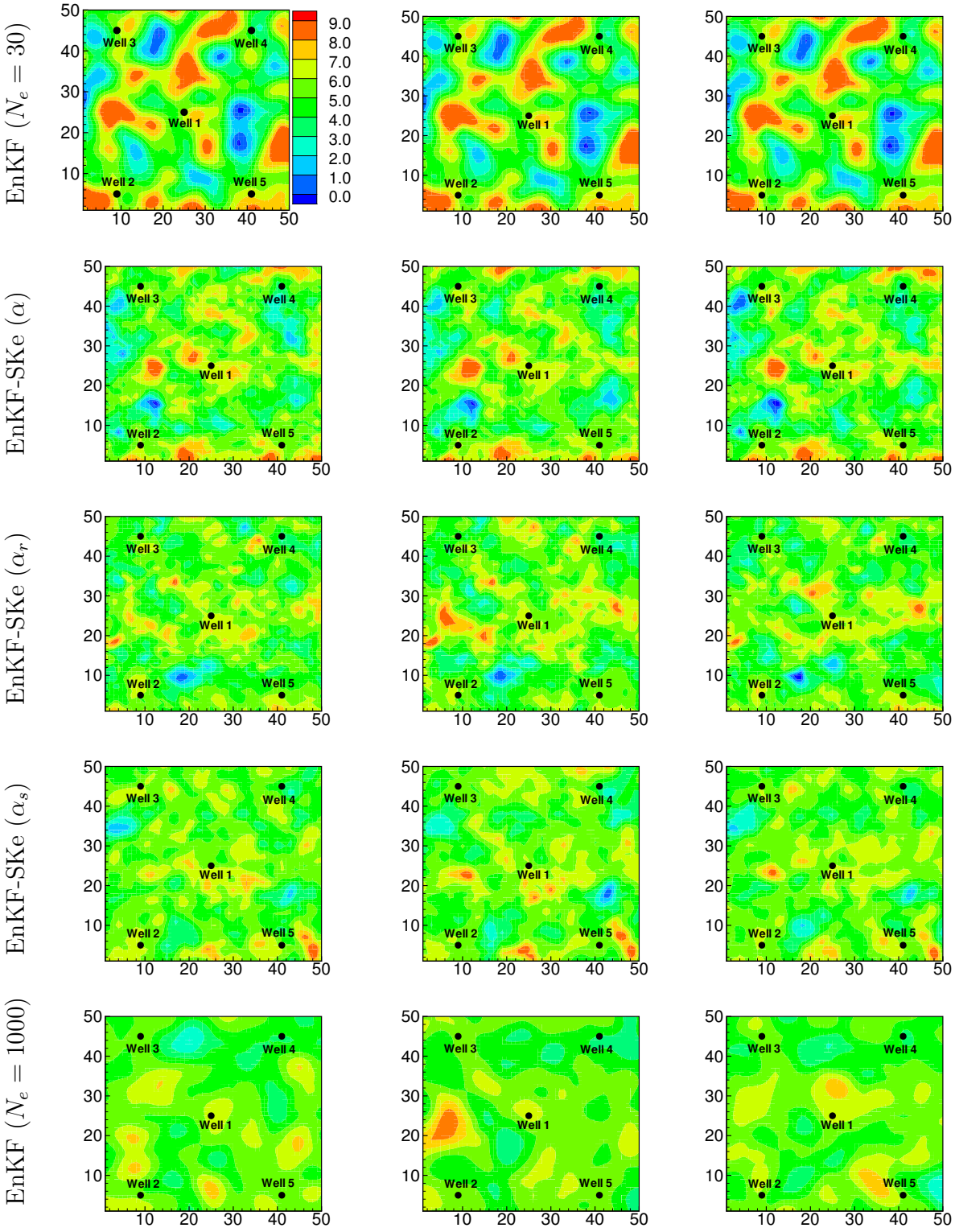
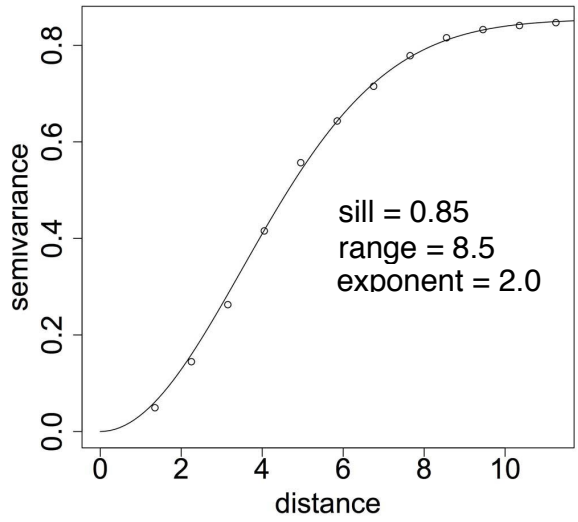
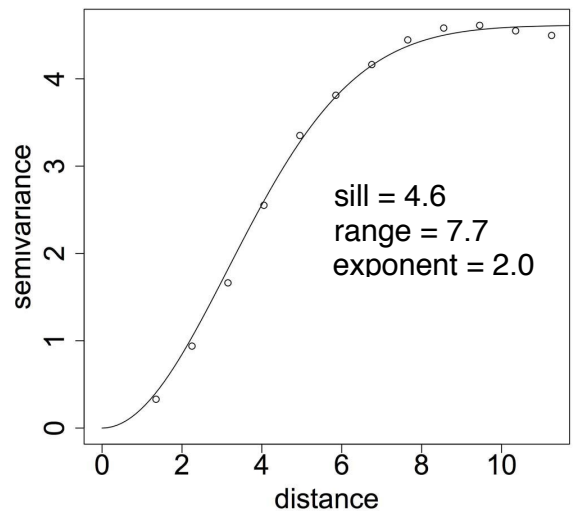


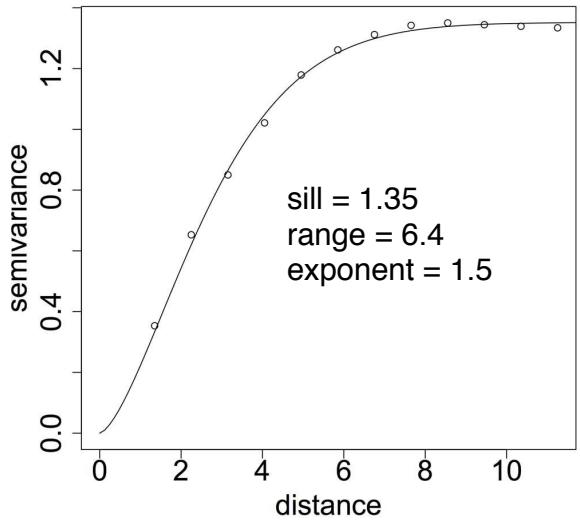
Figure 4.18: Three final updated realizations of log permeability from EnKF and EnKF-SKe.



(a) EnKF ($N_e = 1000$)



(b) EnKF ($N_e = 30$)



(c) EnKF-SKe (α_r)

Figure 4.19: Experimental variograms from final updated realizations of log permeability.

4.6 Chapter summary

In this paper, we proposed an efficient bootstrap version of Anderson's (2007) hierarchical filter and two alternative versions of ensemble Kalman filter with screened Kalman gain (EnKF-SKe) with additional parameters to control the amount of regularization. The three versions of EnKF-SKe were tested on a linear example and a highly nonlinear water flooding reservoir history matching problem. The applications show that the proposed bootstrap methods provided an efficient way of detecting and partially eliminating spurious correlations, and consequently improving the robustness of the Kalman gain. Comparing the three versions of EnKF-SKe, the performance of regularized estimates of screening factor was superior to that of screening factor without regularization in terms of removing spurious correlations. The only additional computational cost incurred by the EnKF-SKe methods is a result of computations of the Kalman gains of the N_B bootstrapped ensembles, but such cost is negligible compared to the cost of running reservoir simulation models. Moreover, even a small number of replicates were generally sufficient for achieving improved history matching results.

Updated realizations from the bootstrapped estimates of the Kalman gain are rougher than realizations from the ensemble of forecasts. One reason for the additional roughness is that all localization methods will cause some reduction in the spatial correlation length (Kepert, 2006). The hierarchical filter of Anderson (2007) and the bootstrapped version described in this paper introduce additional small-amplitude, small-scale roughness through the element-wise multiplication of the Kalman gain by a realization of the screening factors. Localization methods in which the taper function is a smoothly decaying function of distance from the observation location would not inject additional small scale roughness, but lack the generality of bootstrap-based screening method that makes no assumption on spatial continuity.

CHAPTER V

EVALUATION AND ERROR ANALYSIS: KALMAN GAIN REGULARIZATION VERSUS COVARIANCE REGULARIZATION

In petroleum engineering, the ensemble Kalman filter is frequently used for estimating large numbers ($10^5 - 10^6$) of reservoir model parameters and dynamic state variables. The ensemble size is always small compared to the number of variables to be estimated even under the presence of correlations between variables. A small ensemble size introduces statistical sampling error. Under such a situation, we must face the issues of rank deficiency and spurious correlations present in the covariances and the corresponding Kalman gain. A popular method for dealing with these issues is the distance-dependent covariance localization. The concept of localization was originally applied to the covariance matrix. Improved results, however, were also obtained by applying localization on the Kalman gain. In spite of the widespread applications of these two ways of using localization, little in the literature addresses the difference between these two ways of applying localization. This chapter presents a comparison study between the covariance localization and the Kalman gain localization.

The distance-dependent localization is an effective method, but there are some challenges associated with this method (discussed in Chapter 1 and Chapter 4). Thus, in the previous chapter, we present the bootstrap-based screening methods, in which bootstrap resampling method is used to assess the confidence level of each element from the Kalman gain matrix and to filter out the unrealistic correlations from the

Kalman gain. The bootstrap-based screening method was demonstrated to be effective at eliminating unrealistic correlations and easy to implement. In this chapter, we applied the bootstrap-based screening methods on the covariance. A comparison study was also conducted between covariance screening and Kalman gain screening. Therefore, two regularization methods including the distance-dependent localization and the bootstrap-based screening are considered in this work.

The investigations are carried out through two examples: a 1-dimensional linear problem for which the exact solution can be computed and a 2-dimensional highly nonlinear multi-phase reservoir flow problem. In detail, the subjects covered in the investigations include: error evolution in the covariance regularization and Kalman gain regularization, consistency conditions required for the covariance regularization and the applicability of different regularization methods.

5.1 The distance-dependent localization

Distance-dependent localization is the most common method for eliminating spurious correlations. Generally, localization is applied on the covariances by taking the Schur product of covariances with the localization coefficients,

$$K_e^{LC} = C_{yd}^f \circ \beta_{yd} (C_{dd}^f \circ \beta_{dd} + C_D)^{-1} \tag{5.1}$$

where superscript LC stands for localizing covariance, and $\circ$ denotes a Schur or Hadamard product. C_{yd}^f and C_{dd}^f are the two covariances as introduced in Section 2.3. On the other hand, they are also two correlated components of the Kalman gain. To illustrate the relationship, the two covariances are rewritten in terms of linearized sensitivity G

$$\begin{aligned} C_{yd}^f &= E\left[(y-\bar{y})(g(y)-\bar{g}(y))^T\right] \\ &\approx E\left[(y-\bar{y})(y-\bar{y})^T G^T\right] \\ &\approx C_{yy}^f G^T \end{aligned} \tag{5.2}$$

$$
\begin{aligned}
C_{dd}^{f} &= E\left[(g(y)-\bar{g}(y))(g(y)-\bar{g}(y))^{T}\right] \\
&\approx E\left[G(y-\bar{y})(y-\bar{y})^{T}G^{T}\right] \\
&\approx GC_{yy}^{f}G^{T}. \qquad (5.3)
\end{aligned}
$$

Thus, the relationship between the two covariance matrices is $C_{dd}^{f} = GC_{yd}^{f}$ or $C_{dd}^{f} = (C_{yd}^{f})^{T}(C_{yy}^{f})^{-1}C_{yd}^{f}$. This relationship poses consistency requirement for applying covariance localization or other screening algorithms performed on covariances. If we assume that the same consistency condition applies to the localized covariance, then

$$
C_{dd}^{f} \circ \beta_{dd} = G(C_{yd}^{f} \circ \beta_{yd}) \qquad (5.4)
$$

or

$$
C_{dd}^{f} \circ \beta_{dd} = (C_{yd}^{f} \circ \beta_{yd})^{T}(C_{yy}^{f} \circ \beta_{yy})^{-1}(C_{yd}^{f} \circ \beta_{yd}) \ . \qquad (5.5)
$$

For cases, in which, G can be solved efficiently and the cost of computing the full covariance matrix C_{yy}^{f} can also be afforded, there is no need to worry about consistency issue, because the localized C_{yd}^{f} and localized C_{dd}^{f} can be obtained using Eq. 5.2 and Eq. 5.3 with the localized full covariance $C_{yy}^{f} \circ \beta_{yy}$. For most practical applications, however, the problem is nonlinear and high-dimentional, G cannot be computed efficiently and we cannot afford to calculate the full covariance either. Thus, our starting point is the covariances C_{yd}^{f} and C_{dd}^{f}, and the problem is to reduce spurious correlations in these two matrices through construction of taper matrices β_{yd} and β_{dd} that satisfy the consistency conditions (Eq. 5.4 or Eq. 5.5). β_{yd} always can be defined according to the prior model, but it is not trivial to build β_{dd} that is consistent with β_{yd} when the observations are nonlocal. So far, no general methods for defining consistent β_{yd} and β_{dd} are available.

Instead of applying localization to the covariance matrices, an alternative is to apply the localization directly on the Kalman gain by performing a Schur product with β_{yd} (Bergemann and Reich, 2009; Chen and Oliver, 2010; Zhang and Oliver,

2009),

$$K_e^{LK} = \beta_{yd} \circ K_e \ , \tag{5.6}$$

where superscript LK stands for localizing Kalman gain. Kalman gain localization avoids the inconsistency issue, but presents other difficulties as discussed in Chapter 1.

5.2 The bootstrap-based screening

In Chapter 4, we introduce the bootstrap resampling method to compute the variance of an estimator θ that stands for any quantity of interest such as Kalman gain (K_e) or covariance matrices (C_{yd}^f and C_{dd}^f). By using bootstrap method, we avoid evaluating a large number of simulations. The information obtained from bootstrap distribution such as variance or squared variation coefficients are used for calculating screening factor (α). The screening factor α provides an assessment on the accuracy of θ. A small value of α suggests unreliable correlations which should be reduced in magnitude. When θ denotes the Kalman gain, the screening factor α_{ke} is calculated based on N_B replicates of Kalman gain and the screening factor is multiplied to the original estimate of Kalman gain in an element-wise manner:

$$K_e^{SK} = \alpha_{ke} \circ K_e \ , \tag{5.7}$$

where superscript SK stands for screening Kalman gain. Following the screening of the original Kalman gain matrix, the standard updating (or analysis) step is carried out using K_e^{SK}. For later comparison, here we term the EnKF using screened Kalman gain with the short name of EnKF-SKe that is the same as that is used in Chapter 4. Although three ways of defining screening factor are proposed in Chapter 4, only the regularized point-wise estimation is used for obtaining α_{ke} because of its good performance and generality for application.

Above is a review of Kalman gain screening, the covariance screening is similar.

When θ denotes covariances, the screening factors α_{dd} and α_{yd} are calculated respectively for C_{dd}^f and C_{yd}^f using regularized point-wise estimation. Similarly, the screening factors are multiplied to the standard estimates of covariances in an element-wise manner:

$$K_e^{SC} = \alpha_{yd} \circ C_{yd}^f (\alpha_{dd} \circ C_{dd}^f + C_D)^{-1} \ , \tag{5.8}$$

where the superscript SC denotes screening covariance. The consistency between α_{dd} and α_{yd} is an issue to be discussed in later examples. This method is denoted as EnKF-SCov.

5.3 1D linear problem

In this section, we investigate several approaches to the reduction of spurious correlations in data assimilation on a 1D correlated random field $X = \{x_1, x_2, \ldots, x_{100}\}$ with prior mean 0 and exponential covariance as shown in Eq. 5.9. The covariance function has an exponent of 1.5 and the range r varies in two testing scenarios,

$$C(h) = \exp[-3.(|h|/r)^{1.5}] \tag{5.9}$$

where h is the distance between two points.

In this example, the observations are directly the measurements of state variables, thus the sensitivity G is a matrix that contains 1 at a data location and 0 everywhere else. The true Kalman gain for this problem is computed using the known sensitivity matrix and prior covariance (Eq. 5.9). To compare the different ensemble-based estimates of Kalman gain with and without screening or localization, an initial ensemble of independent, unconditional realizations are drawn from the same distribution as the prior for X.

The distance-dependent localization coefficients matrix β_{yy} used in the algorithms of covariance localization and Kalman gain localization is defined using the following

equation (Furrer and Bengtsson, 2007)

$$\beta(h) = \frac{1}{1 + (1 + \frac{1}{C(h)^2})/N_e} \ . \tag{5.10}$$

A consistent pair of β_{yd} and β_{dd} is always easy to build for the linear local measurement problem. Because of the simplicity of the sensitivity, β_{yd} and β_{dd} are simply the block matrices extracted from β_{yy} by $\beta_{yd} = \beta_{yy}G^T$ and $\beta_{dd} = G\beta_{yd}$. For bootstrap-based screening covariance methods, the consistency issue is addressed in the single observation test.

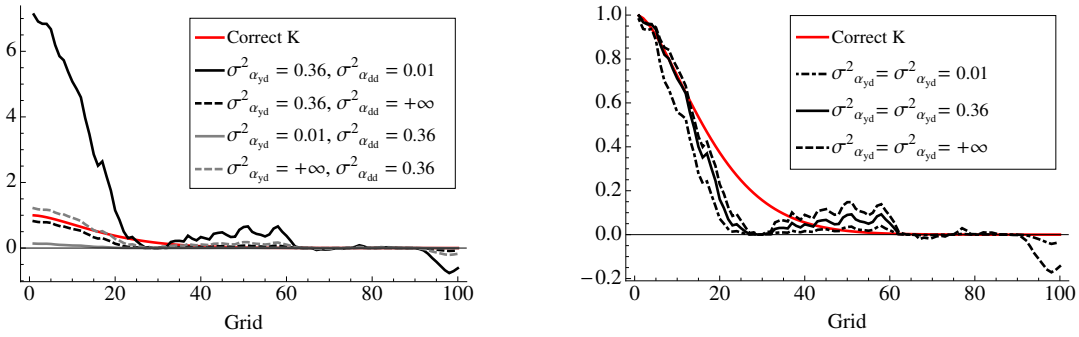
In all the following testing scenarios, the number of bootstrapped ensembles, $N_B = 100$, and all the screening factors (α_{dd}, α_{yd}, and α_{ke}) are calculated based on the same set of bootstrap ensembles. The estimates of the Kalman gain using different localizing or screening algorithms: K_e^{LC}, K_e^{LK}, K_e^{SK}, and K_e^{SC} are compared with the true Kalman gain in the following two tests.

5.3.1 Single observation

The range, $r = 40$ and the ensemble size, $N_e = 30$ are used in this example. A single measurement with additive Gaussian noise (mean 0 and standard error of 0.05) is made at the first gridblock. Hence, the measurement error covariance C_D is a scalar and has a value of 0.0025.

For this linear local measurement problem, the screening covariance satisfies the consistency condition ($\alpha_{dd} = G\alpha_{yd}$) as long as the same value of σ_α^2 is used for calculating both α_{yd} and α_{dd}. Fig. 5.1 shows estimates of Kalman gain from different combinations of $\sigma^2_{\alpha_{yd}}$ and $\sigma^2_{\alpha_{dd}}$. For the cases of $\sigma^2_{\alpha_{yd}} \neq \sigma^2_{\alpha_{dd}}$, the estimate of Kalman gain is not correct at the data location because the consistency condition is violated. Fig. 5.1 (b) shows that when the same value of σ_α^2 is used for both covariance matrices, the estimates are good at the data location.

Fig. 5.2 compares the standard estimate of the Kalman gain to the estimates of



(a) Inconsistent pair (b) Consistent pair

Figure 5.1: Influence and relations of $\sigma^2_{\alpha_{yd}}$ and $\sigma^2_{\alpha_{dd}}$.

Kalman gain with screening or distance-based localization applied. Screening and localization both reduce the magnitude of spurious correlations present in the standard estimate of Kalman gain. In this test, since $\beta_{dd} = 1.0$, the estimate of Kalman gain from covariance localization (K_e^{LC}) is exactly the same as the estimate from Kalman gain localization (K_e^{LK}). The distance-based localization methods completely remove the unreal non-zero values beyond grid 55 due to the zero values of coefficients β_{yd} for grids that are located outside the correlation length. Although the bootstrap-based screening algorithms did not completely eliminate the spurious correlations, the magnitudes of the spurious correlations were reduced. Screening Kalman gain performs slightly better than screening covariance in this example.

In order to reduce the influence of randomness on results, this 1D single observation testing case was evaluated 100 times, each time with different randomly generated initial ensemble. The root mean squared error (RMSE) for different estimates of Kalman gain are calculated using

$$\text{RMSE}_i = \sqrt{\frac{\sum_{n=1}^{100}(K_{ei,n} - K_i)^2}{100}}\ , \tag{5.11}$$

where n is the index of the trial, and i is the index of the element in the Kalman gain. K_e denotes the ensemble-based estimate of Kalman gain from different methods,

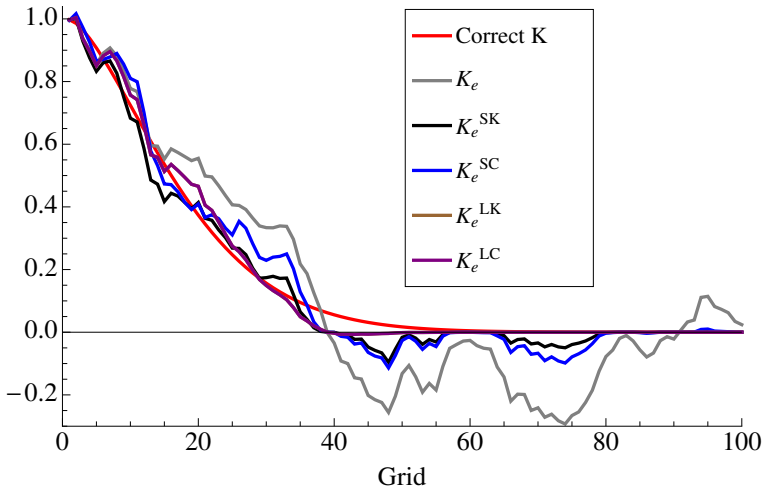


Figure 5.2: The estimates of Kalman gain.

and K_i is the true Kalman gain. Fig. 5.3 shows that the distance-based localization methods have the lowest RMSE. Standard EnKF without localization or screening resulted in the highest RMSE values in the region that is distant from data location. Screening the Kalman gain results in smaller RMSE than screening the covariance. Both screening methods, however, have high RMSE values around x_{20}. The true Kalman gain or C_{yd} is fairly large at x_{20}, and the variability is also large. The screening methods sometimes reduce the magnitude in that region more than necessary, so the results from the screening methods are slightly worse on average than the standard estimate in that region.

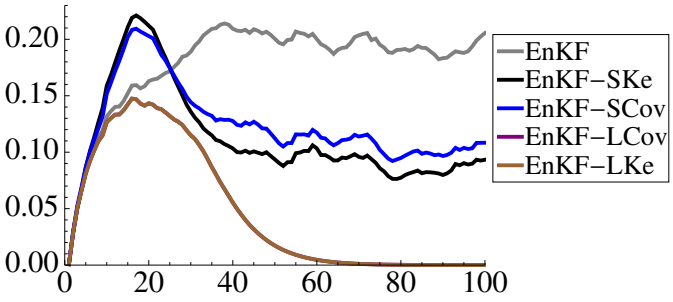
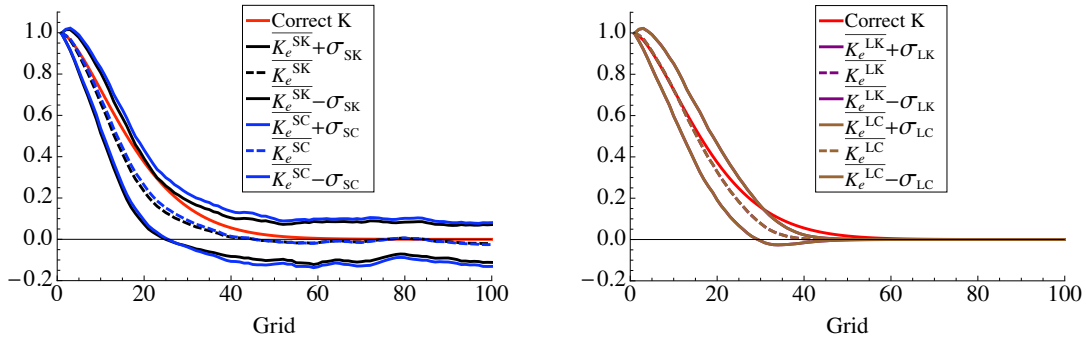


Figure 5.3: Root mean squared error (RMSE) of Kalman gain estimates.

Fig. 5.4 shows the mean estimate of Kalman gain along with one standard deviation from screening Kalman gain (K_e^{SK}), screening covariance (K_e^{SC}), localizing covariance (K_e^{LC}), and localizing Kalman gain (K_e^{LK}). Localization and screening result in some bias shown in the expected Kalman gain curve compared to the correct Kalman gain. Due to the effect of distance-based localization, the estimate of Kalman gain obtained from localization methods only shows uncertainty within the correlation length, beyond which the estimated values exactly go to zero. Screening the covariance shows slightly larger standard deviation than screening the Kalman gain.



(a) Screening Kalman gain (black), screening covariance (blue)

(b) Localizing Kalman gain (brown), localizing covariance (purple)

Figure 5.4: Mean estimate of Kalman gain with one standard deviation.

To understand why screening the covariance results in greater variability than screening the Kalman gain, we look at the propagations of uncertainty (or error) in these two algorithms. For any function, $f(x, y)$, the linearized approximation to the variance (or error) of f that is propagated from the variances in the scalar variables x and y can be computed as

$$\sigma_f^2 = (\frac{\partial f}{\partial x})^2\sigma_x^2 + (\frac{\partial f}{\partial y})^2\sigma_y^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\sigma_x\sigma_y\rho_{xy} \ , \tag{5.12}$$

where ρ_{xy} is the correlation coefficient between x and y, σ_x and σ_y are standard deviations (or standard errors) associated with variables x and y respectively. Using this

equation, we can estimate the variances within K_e^{SK} and K_e^{SC} due to the variability in the screening coefficients.

For any element i in the example with a single observation,

$$K_{e,i}^{SK} = \frac{C_{yd,i}}{C_{dd} + C_D}\alpha_{ke,i}$$

$$K_{e,i}^{SC} = \frac{C_{yd,i}\alpha_{yd,i}}{C_{dd}\alpha_{dd} + C_D} \ .$$

The terms C_D, $C_{yd,i}$, C_{dd} are common in the above two equations, thus we ignore their contributions. $K_{e,i}^{SK}$ is a function of $\alpha_{ke,i}$, and $K_{e,i}^{SC}$ is a function of both $\alpha_{yd,i}$ and α_{dd}. Following Eq. 5.12, the relative variance $\tilde{\sigma}_f^2 = \sigma_f^2/f^2$ for $K_{e,i}^{SK}$ is

$$\tilde{\sigma}_{sk,i}^2 = \frac{\sigma_{\alpha_{ke,i}}^2}{\alpha_{ke,i}^2} \ . \tag{5.13}$$

Similarly, the relative variance for $K_{e,i}^{SC}$ is

$$\tilde{\sigma}_{sc,i}^2 = \frac{\sigma_{\alpha_{yd,i}}^2}{\alpha_{yd,i}^2} + \frac{\sigma_{\alpha_{dd}}^2}{(\alpha_{dd} + \frac{C_D}{C_{dd}})^2} - 2\rho\left(\frac{\sigma_{\alpha_{yd,i}}}{\alpha_{yd,i}}\right)\left(\frac{\sigma_{\alpha_{dd}}}{\alpha_{dd} + \frac{C_D}{C_{dd}}}\right) \ . \tag{5.14}$$

As mentioned previously, the variance of K_e is not included here, since it is common for both K_e^{SK} and K_e^{SC}. Then, the relative variance of K_e^{SK} is only determined by variance of α_{ke} as shown in Eq. 5.13, while the relative variance of K_e^{SC} is determined by the sum of three error terms (Eq. 5.14). $\sigma_{\alpha_{dd}}^2$ is definitely larger than zero, $\sigma_{\alpha_{yd}}^2$ and $\sigma_{\alpha_{ke}}^2$ are comparable in scale as shown in Fig. 5.5.

In the neighborhood of data location ($i = 1$), $\sigma_{\alpha_{yd}}^2$ is significantly larger than $\sigma_{\alpha_{ke}}^2$. The reason for $\tilde{\sigma}_{sc,i}^2 \approx 0$ at data location as shown in Fig. 5.4 (a) is that $\alpha_{yd,i} = \alpha_{dd}$, $\sigma_{\alpha_{dd}} = \sigma_{\alpha_{yd,i}}$, and the correlation coefficient $\rho = 1$ since $\alpha_{yd,i}$ and α_{dd} are perfectly correlated. Moreover, $\frac{C_D}{C_{dd}}$ is typically a very small value, which means the absolute value of the negative term approximately equals to the positive terms. The reason for $\tilde{\sigma}_{sk,i}^2 \approx 0$ at data location is that the Kalman gain value at data location is always approximately 1 as $K_{e,i} = \frac{C_{dd}}{C_{dd}+C_D}$ at the data location, which does not change with ensemble. As i increases, however, the standard deviation of $K_{e,i}^{SC}$

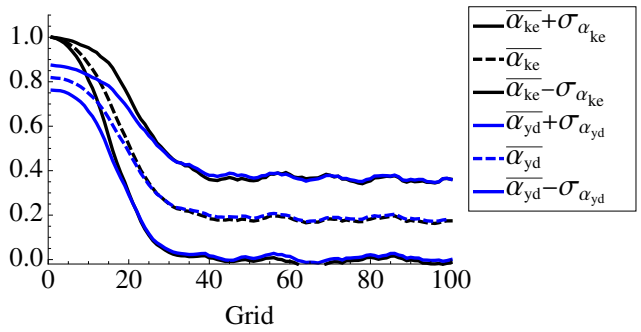


Figure 5.5: Mean estimates with one standard deviation for α_{ke} and α_{yd} based on 100 trials.

is constantly larger than $K_{e,i}^{SK}$. This is because the correlation between data and model parameter becomes smaller as we move away from the data location, which leads to the decrease of the correlation coefficient between $\alpha_{yd,i}$ and α_{dd}. As $\rho \to 0$, $\tilde{\sigma}_{sc,i}^2 \to \frac{\sigma_{\alpha_{yd,i}}^2}{\alpha_{yd,i}^2} + \frac{\sigma_{\alpha_{dd}}^2}{(\alpha_{dd}+\frac{C_D}{C_{dd}})^2}$, and the positive term $\frac{\sigma_{\alpha_{dd}}^2}{(\alpha_{dd}+\frac{C_D}{C_{dd}})^2}$ is the main contribution for the constant difference between $\tilde{\sigma}_{sc,i}^2$ and $\tilde{\sigma}_{sk,i}^2$ in the region beyond correlation length.

For multiple data, $\tilde{\sigma}_{sk,i}^2$ still only depends on $\sigma_{\alpha_{ke,i}}^2$, but $\tilde{\sigma}_{sc,i}^2$ is a function of $N_d(N_d+3)/2$ variables (N_d is the number of data). Estimation of the Kalman gain from screening the covariance is more error sensitive than screening the Kalman gain directly. This conclusion is also true for covariance localization and Kalman gain localization. Although for this 1D single measurement test, covariance localization is exactly the same as Kalman gain localization, it is not generally the case for multiple data. In addition, while the distance-based localization methods performed better than the screening algorithms for this test, it should be noted that the localization coefficients were calculated based on the true prior covariance, in a problem which was ideal for application of distance-based localization. For more general real problems, these ideal conditions for localization do not apply, and distance-based localization can be difficult.

5.3.2 Multiple observations

The screening and localization algorithms all performed well in the test involving a single measurement. In the present scenario, we consider multiple spatially correlated observations for the same 1-dimensional model. The main objective is to investigate the impact of influence between different data on the estimates of Kalman gain from EnKF with screening or localization. Five measurements at x_1, x_{25}, x_{50}, x_{75} and x_{100} are used for data assimilation. The measurement error covariance C_D is a diagonal matrix with the same diagonal values of 0.0025. The ensemble size is 30. The range of prior covariance is 100, so, the 5 measurements are spatially correlated with each other. The true data covariance C_{dd}^f is

$$\begin{pmatrix} 1. & 0.702769 & 0.357364 & 0.148122 & 0.0520728 \\ 0.702769 & 1. & 0.687289 & 0.346227 & 0.142479 \\ 0.357364 & 0.687289 & 1. & 0.687289 & 0.346227 \\ 0.148122 & 0.346227 & 0.687289 & 1. & 0.687289 \\ 0.0520728 & 0.142479 & 0.346227 & 0.687289 & 1. \end{pmatrix}.$$

The true Kalman gain matrix consists of five columns that correspond to the five measurements, respectively. These five columns are plotted together and are shown in Fig. 5.6. The results from this test can be understood by analyzing the result of any one column from the Kalman gain matrix. Therefore, only the results of column 1 will be presented here.

Fig. 5.7 shows the first column of ensemble-based estimates of Kalman gain from different methods. Screening Kalman gain (K_e^{SK}) and localizing Kalman gain (K_e^{LK}) only reduce the magnitude of spurious correlations, and do not change the sign of correlations, screening covariance and localizing covariance, however, can change the structure of Kalman gain, for example, the negative values are changed to be positive values between grid 80 to grid 100. It is highly possible that screening/localizing

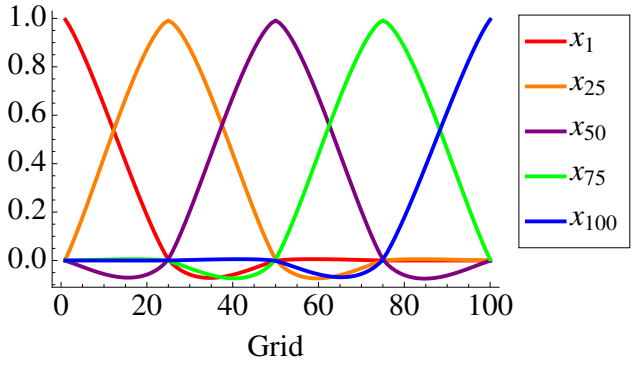


Figure 5.6: True Kalman gain consisting of 5 columns corresponding to data at: x_1, x_{25}, x_{50}, x_{75} and x_{100}.

covariance can introduce spurious correlations of larger magnitude. The advantage of screening Kalman gain over localizing Kalman gain is also indicated in Fig. 5.7. Around grid 40, there are evidently spurious correlations present in the standard estimate of Kalman gain (K_e) of opposite sign to the true correlations shown in correct Kalman gain. The screening Kalman gain (K_e^{SK}) decreased the magnitudes, but the localizing Kalman gain (K_e^{LK}) had no effect. As the correlation length is 100 for this test, the values of distance-based localization coefficients around grid 40 are high, between 0.85 and 0.95, therefore, localization can not eliminate the spurious correlations shown in the high correlation region.

In order to obtain reliable statistical conclusions, the test was repeated 100 times with different initial ensembles. Fig. 5.8 shows the mean estimates of Kalman gain with one standard deviation. The estimate from screening the Kalman gain has significantly smaller standard deviation than the estimate from screening the covariance. Similarly, the result from localizing the Kalman gain has smaller standard deviation than that from localizing the covariance, especially between grid 80 and grid 100. Fig. 5.9 shows the root mean squared error (RMSE) of estimates of Kalman gain from the different methods. Only in the region (between grid 1 and grid 20) where

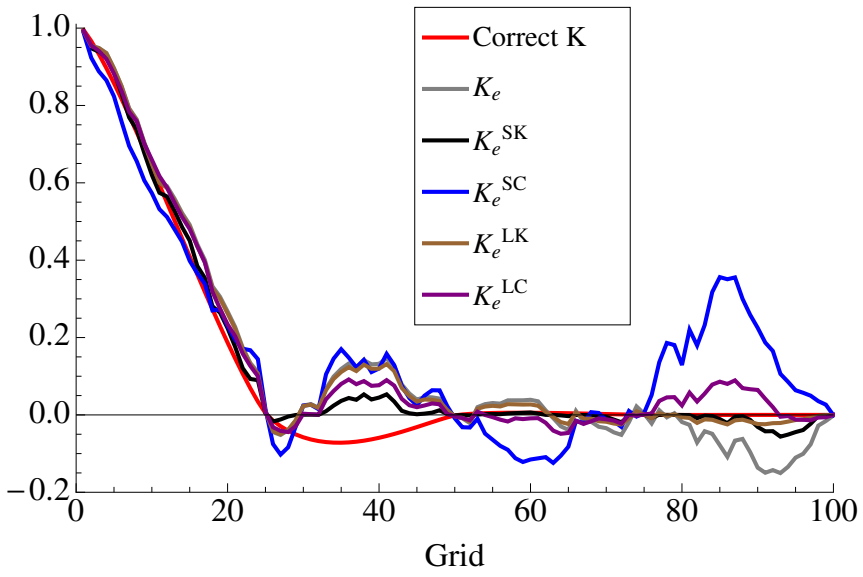
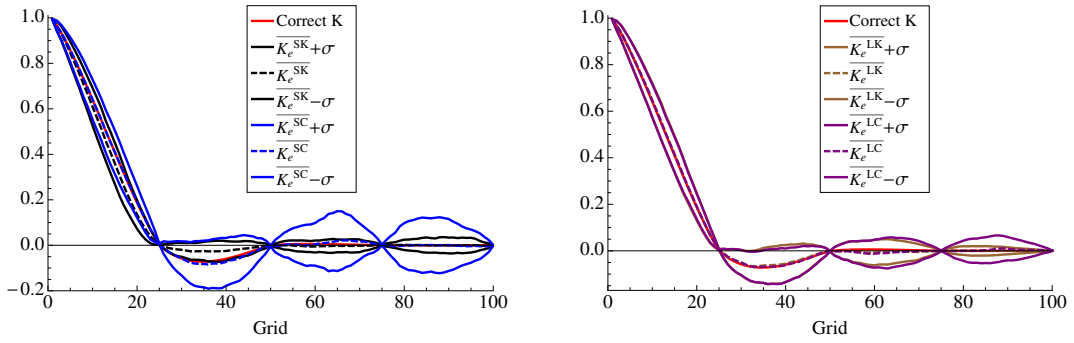


Figure 5.7: The estimates of Kalman gain.

data and model parameters are most highly correlated, screening/localizing covariance results in smaller RMSE values than screening/localizing Kalman gain. In the rest of the region, screening/localizing covariance results in higher RMSE than screening/localizing Kalman gain, which is consistent with the conclusion from the error analysis that the canceling effect of negative crossing terms vanishes as the correlation coefficients get close to zero. The RMSE values are nearly zero at data locations, because the measurements are directly of the model variables. The rows of C_{dd} are identical to the rows of C_{yd} corresponding to the 5 data locations, thus the values of the Kalman gain at the 5 data locations obtained from $C_{dd}(C_{dd} + C_D)^{-1}$, which is approximately an identity matrix when the magnitudes of the entries in C_D are very small. Thus, regardless of the values in covariances C_{dd} or C_{yd}, the values of the 1st column of Kalman gain matrix at 5 data locations are approximately 1, 0, 0, 0, 0 respectively. If we neglect the sharp decrease in RMSE values at data locations, the RMSE values from the screening Kalman gain become smaller as the distance from x_1 increases. On the other hand, screening covariance shows an opposite trend.



(a) Screening Kalman gain (black), screening covariance (blue)

(b) Localizing Kalman gain (brown), localizing covariance (purple)

Figure 5.8: Mean estimate of Kalman gain with one standard deviation.

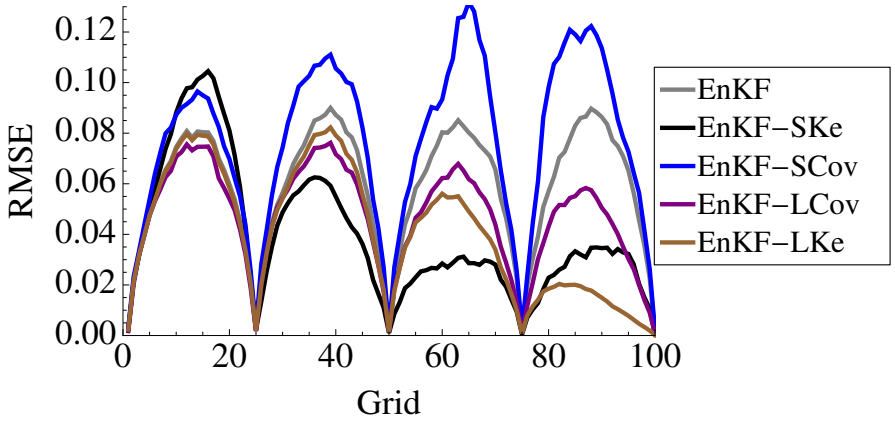
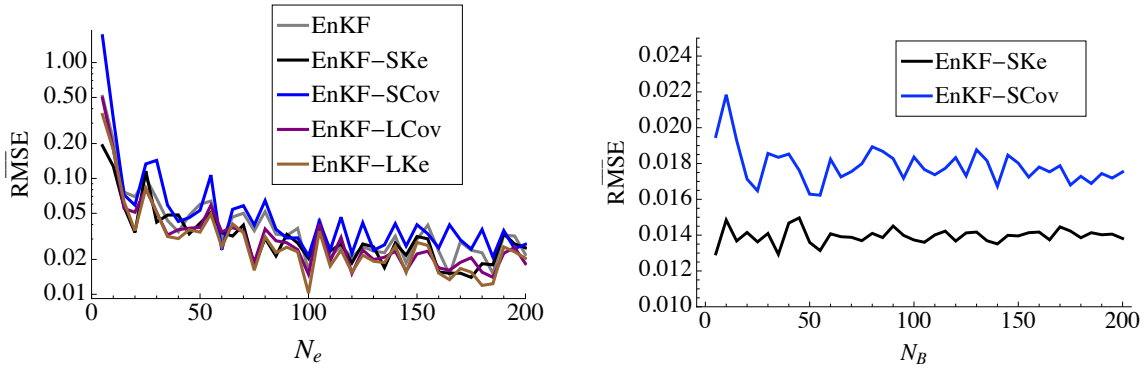


Figure 5.9: Root mean squared error of Kalman gain estimates.

In order to see to what extent the screening/localizing algorithms are influenced by the ensemble size (N_e) or the number of bootstrap samples (N_B), a simple sensitivity study was performed. The criterion for quantifying the influence is the mean RMSE computed using Eq. 5.15. Fig. 5.10 (a) shows the logarithm of $\overline{\text{RMSE}}$ versus ensemble size N_e. The tested ensemble size is gradually increased from 5 to 200 by a step size of 5. When $N_e = 5$, screening Kalman gain has smallest $\overline{\text{RMSE}}$ value. When $N_e \geq 100$, $\overline{\text{RMSE}}$ is small for all methods. The $\overline{\text{RMSE}}$ of the estimate of the Kalman gain is smaller from screening the Kalman gain than that obtained from screening the covariance for all different ensemble sizes. Localizing the Kalman gain also results in slightly smaller $\overline{\text{RMSE}}$ than localizing the covariance. Fig. 5.10 (b) shows how the screening methods are influenced by the number of bootstrap samples for N_B from 5 to 200. Screening Kalman gain is not very sensitive to N_B, while screening covariance seems to be influenced by N_B, only when N_B is very small. The $\overline{\text{RMSE}}$ values from screening covariance are consistently higher than those from screening Kalman gain.

$$\overline{\text{RMSE}} = \sqrt{\frac{\sum_{i=1}^{N_y}(K_{e,i} - K_i)^2}{N_y}} \tag{5.15}$$



(a) N_e influence ($N_B = 100$ for screening methods)

(b) N_B influence ($N_e = 200$)

Figure 5.10: Sensitivity study.

In the 1D linear problem with single/multiple observations, we were able to use

the true prior covariance and true range in the localization function. In sequential data assimilation, this may not be the case, since the prior covariance changes with time due to the assimilated data at each timestep (Chen and Oliver, 2009). Thus, in the next section, a comparison study is carried on a sequential data assimilation on a nonlinear 2D reservoir flow model.

5.4 2D highly nonlinear problem

5.4.1 Reference model

Reference data for evaluation of the methods are generated from a reference reservoir model that is 100×100 with individual gridblock dimensions of 30 ft $\times$ 30 ft $\times$ 20 ft. The wells are drilled in a repeat five-spot water flooding pattern. There are 15 producers and 15 injectors in the field. Porosity is 0.20 for all gridblocks. The only uncertain model parameter in this problem is log permeability at each grid block. The reference log permeability field is generated using an isotropic exponential variogram model with a practical range of 10 gridblocks, mean of 3.5, and standard deviation of 1.0. Fig. 5.11 shows the reference log permeability field with black circles denoting the locations of production wells and triangles denoting the locations of water injection wells.

The producers are controlled by fixed bottom hole pressure with maximum oil production rate as the secondary constraint. The injectors are controlled by fixed water injection rate with maximum bottom hole pressure as the secondary constraint. The reference reservoir model is produced for a total of 520 days. Fig. 5.12 shows the production data profiles from the reference model which include the oil and water production rates for the producers as well as the bottom hole pressure of the injectors.

5.4.2 Test setup

The total production period for the reservoir model is 520 days. The time between day 0 and day 250 is treated as the production history and the period from day

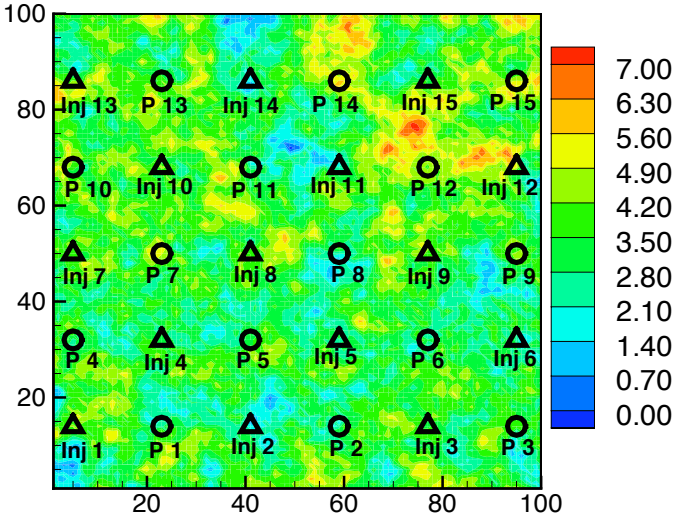


Figure 5.11: The reference log permeability field.

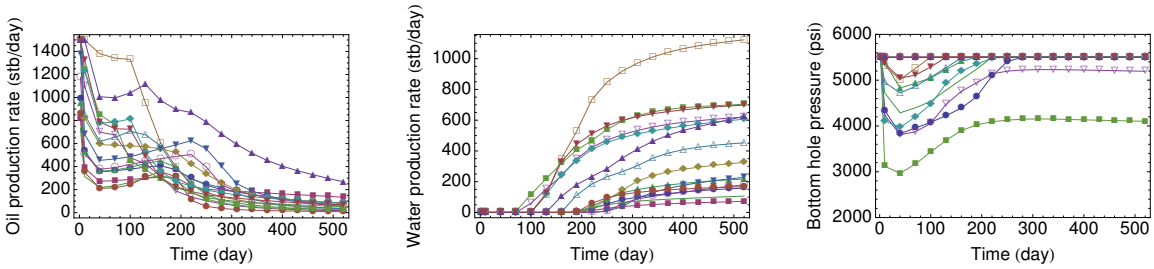


Figure 5.12: The production profiles from the reference model (different curves denote different wells).

251 to day 520 is considered as the prediction period. Water injection in the field started from day 0 and continued until the end of the production period (520 days). The oil and water production rate data from each producer and the bottom hole pressure data from each injector are used as observations during data assimilation. The observations are taken at day 10, and every 60 days thereafter until day 250. Thus, there are a total of 5 data assimilation time steps and 45 production data at each assimilation step. The measurement noise for the injector bottom hole pressure and oil production rate are assumed to have a mean of 0 and the standard deviation of measurement error is assumed to be 1% of the actual observation value. The standard deviation of measurement error for water production rate data is assumed to be 2% of the actual observation value.

In order to verify the ability of different screening and localizing algorithms for eliminating spurious correlations, a small ensemble containing 30 members was used, which is likely to result in significant sampling errors. There are a total of 225 data to be assimilated during 5 data assimilation steps using this small ensemble. Log permeability, pressure, and water saturation are included into the state vector. Thus, the state vector for each ensemble member contains a total of 30,000 model parameters and state variables.

For this high-dimensional data assimilation problem, the measurements are non-local. A consistent covariance localization involves computing the full covariance matrix. To avoid the intensive cost of computing the full covariance matrix, Chen and Oliver (2009) proposed an approximate form for constructing β_{dd} by replacing β_{yy} with an identity matrix,

$$\beta_{dd} = \beta_{yd}^T \beta_{yd} \ . \tag{5.16}$$

The authors also showed that acceptable results were obtained by using the proposed approximation. Therefore, in this test, Eq. 5.16 is used for computing β_{dd}, and β_{yd} is built using Eq. 5.10 with a range of 25 gridblocks that is determined based on the

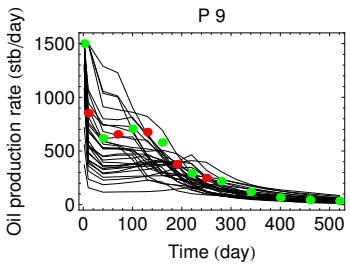
correlation length of the prior log permeability field together with the sensitivity and well pattern information. The same localization function is used for all three types of data. Both log permeability and dynamic state variables (including water saturation and pressure) are updated with localization.

For the two bootstrap-based screening algorithms, $N_B = 50$, $\sigma_\alpha^2 = 0.36$, and the same random seed is used during bootstrapping. Two more cases were also evaluated for comparison, including the standard EnKF with a small ensemble size of 30 and that with a fairly large ensemble size of 2000. For this complex flow model, we do not know the exact Kalman gain so the estimate of Kalman gain from EnKF with $N_e = 2000$ is used for comparison with the estimates from the other methods.

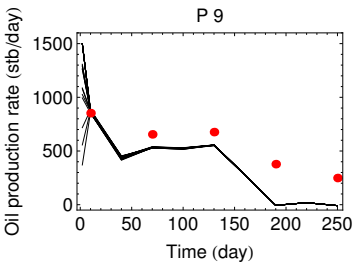
5.4.3 Match production data

The variability represented by the initial ensemble is able to cover most of the observations from the reference model. As an example, Fig. 5.13 (a) shows the predictions of oil production rate of producer P 9 from the initial ensemble prior to assimilating any observations. After assimilating 45 data at the first data assimilation step, the updated ensemble from the standard EnKF loses nearly all the ensemble variability as shown in Fig. 5.13 (b) and 5.13(c), which illustrate the necessity of applying screening or localization algorithms.

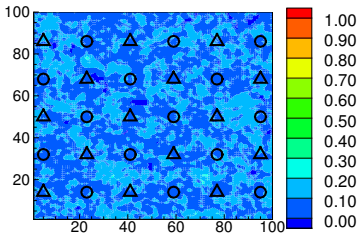
Once the entire data assimilation process is complete, the final updated ensemble of log permeability was evaluated from the beginning (day 0) up to the end of the production period (day 520) using a commercial reservoir simulator (Schlumberger, 2007). Fig. 5.14 shows the predictions of different production data for three wells obtained by rerunning the final updated log permeability fields from day 0 to day 520. The standard EnKF without covariance/Kalman gain regularization is not able to match the production profiles because of the ensemble collapse that was observed



(a) Initial



(b) During data assimilation



(c) The grid-based ensemble standard deviation of log permeability after 1st data assimilation

Figure 5.13: The loss of ensemble variability for standard EnKF with an ensemble size of 30 (red dots denote observations used for data assimilation, green dots denote observations only for comparison, black curves denote ensemble outputs in subfigures (a) and (b)).

at early data assimilation timesteps. The EnKF with screening covariance (EnKF-SCov) shows larger ensemble variability, but does not match data well. The remaining three methods (EnKF-SKe, EnKF-LKe, and EnKF-LCov) have comparatively good matches to the reference production data from the reference model. In order to make a quantitative evaluation of the data matches from different methods, two evaluation criteria are defined including the average root mean squared error $\hat{e}_d$

$$\hat{e}_d = \frac{1}{N_t N_w} \sum_{t=1}^{N_t} \sum_{w=1}^{N_w} \sqrt{\frac{1}{N_e} \sum_{i=1}^{N_e} (d_{t,w}^{obs} - d_{t,w,i})^2} \ , \tag{5.17}$$

and the average prediction spread $\hat{\sigma}_d$

$$\hat{\sigma}_d = \frac{1}{N_t N_w} \sum_{t=1}^{N_t} \sum_{w=1}^{N_w} \sqrt{\frac{1}{N_e} \sum_{i=1}^{N_e} (d_{t,w,i} - \langle d_{t,w} \rangle)^2} \ , \tag{5.18}$$

where N_e is the number of ensemble members, N_t is the number of data records, N_w is the number of wells for the same type of data, $d_{t,w}^{obs}$ denotes observation, $d_{t,w,i}$ denotes predicted data, and $\langle d_{t,w} \rangle$ denotes the mean of ensemble prediction. Table 5.1 shows the $\hat{e}_d$ versus $\hat{\sigma}_d$ for three types of data. The standard EnKF results in the largest $\hat{e}_d$ and smallest $\hat{\sigma}_d$, which is the worst case followed by EnKF with screening covariance (EnKF-SCov). The other three methods result in similar $\hat{e}_d$ for all data types, while

EnKF-SKe has relatively smaller $\hat{\sigma}_d$ than EnKF-LKe and EnKF-LCov.

Table 5.1: The average error versus average spread ($\hat{e}_d/\hat{\sigma}_d$) of the predictions of three types of data: OPR (Oil Production Rate), WPR (Water Production Rate) and BHP (Bottom Hole Pressure).

	OPR (stb/day)	WPR (stb/day)	BHP (psi)
EnKF (N_e=30)	170/0.6	149/0.7	565/2.2
EnKF-SCov	190/33	144/24	220/32
EnKF-SKe	59/19	42/16	144/51
EnKF-LKe	68/46	43/29	89/75
EnKF-LCov	97/39	42/27	92/67

5.4.4 Estimates of model parameter (log permeability)

Fig. 5.15 shows the final estimates (ensemble mean) of log permeability obtained from different methods. The estimate of log permeability from the standard EnKF (N_e = 30) shows extremely high and low values. For EnKF-SCov, the estimates also suffer from the overshooting issue and are highly discontinuous. The other three methods result in better estimates of log permeability as the magnitudes of estimates are similar to those of the reference model. There are, however, some artificial phenomena appearing in the estimate from EnKF-LCov as the log permeability values around the production well P 2 ($x = 59$, $y = 14$) appear significantly different in magnitudes from the region outside the neighborhood of well P 2. The inconsistency issue might be responsible for this behavior. Table 5.2 shows the quantitative evaluation of the estimates of log permeability obtained from different methods. Spatial mean is the average value of log permeability over all gridblocks. The true spatial mean has a value of 3.5, so two methods including EnKF-SKe and EnKF ($N_e = 2000$) provide accurate estimates of the spatial mean. The ensemble STD is the average of grid-based ensemble standard deviation of log permeability. Compared to the ensemble STD obtained from EnKF ($N_e = 2000$), the ensemble STD from EnKF-LKe is high, while the values from EnKF ($N_e = 30$) and EnKF-SCov are low. RMSE is the average

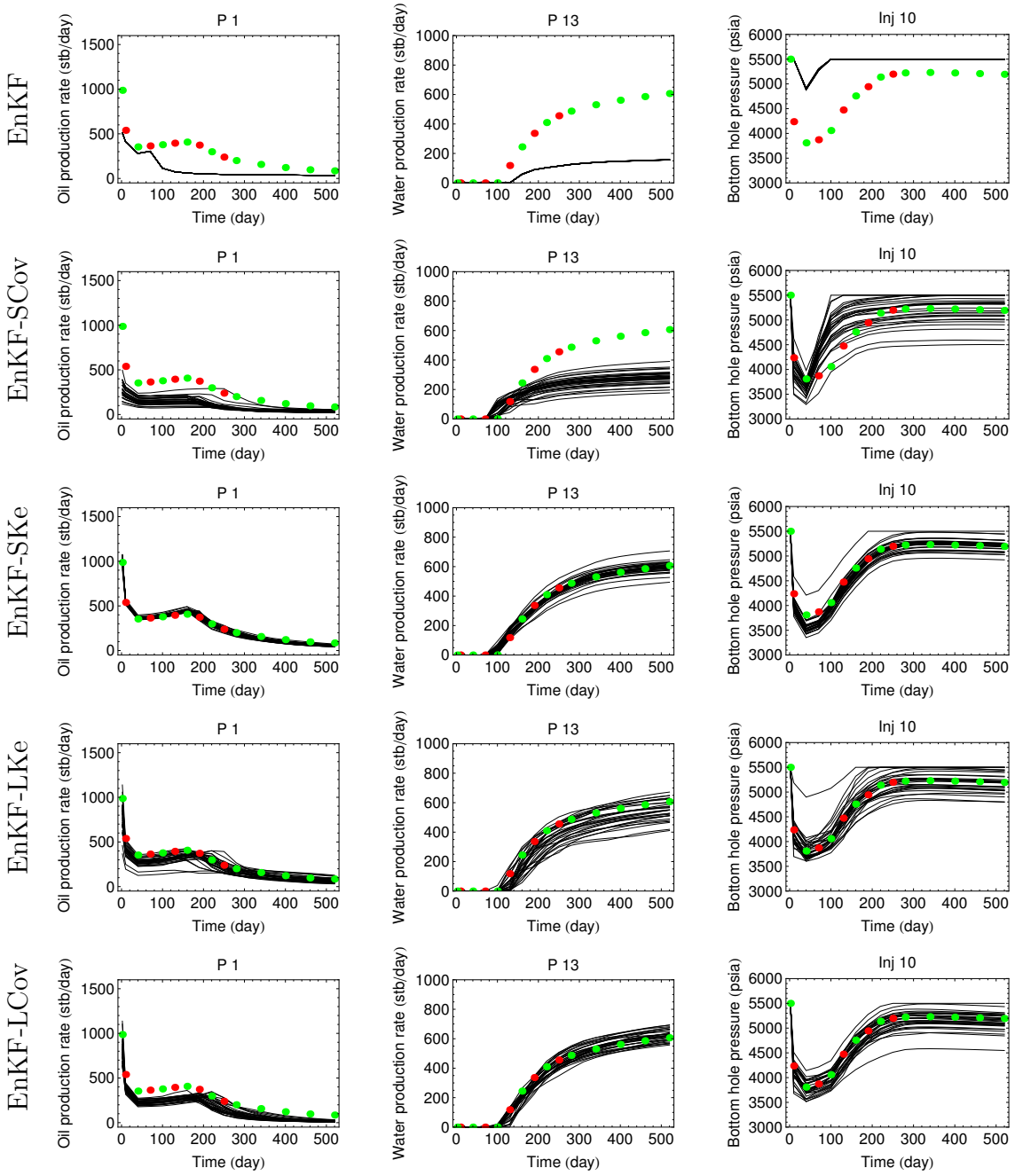


Figure 5.14: Ensemble predictions based on final estimated log permeability fields for wells P 1, P 13, Inj 10: observations used for data assimilation (red dots), observations only for comparison (green dots), ensemble outputs (black lines).

grid-based root mean squared error of estimates of log permeability. The RMSE values from EnKF-SKe and EnKF-LCov are similar to the RMSE values obtained from EnKF ($N_e = 2000$).

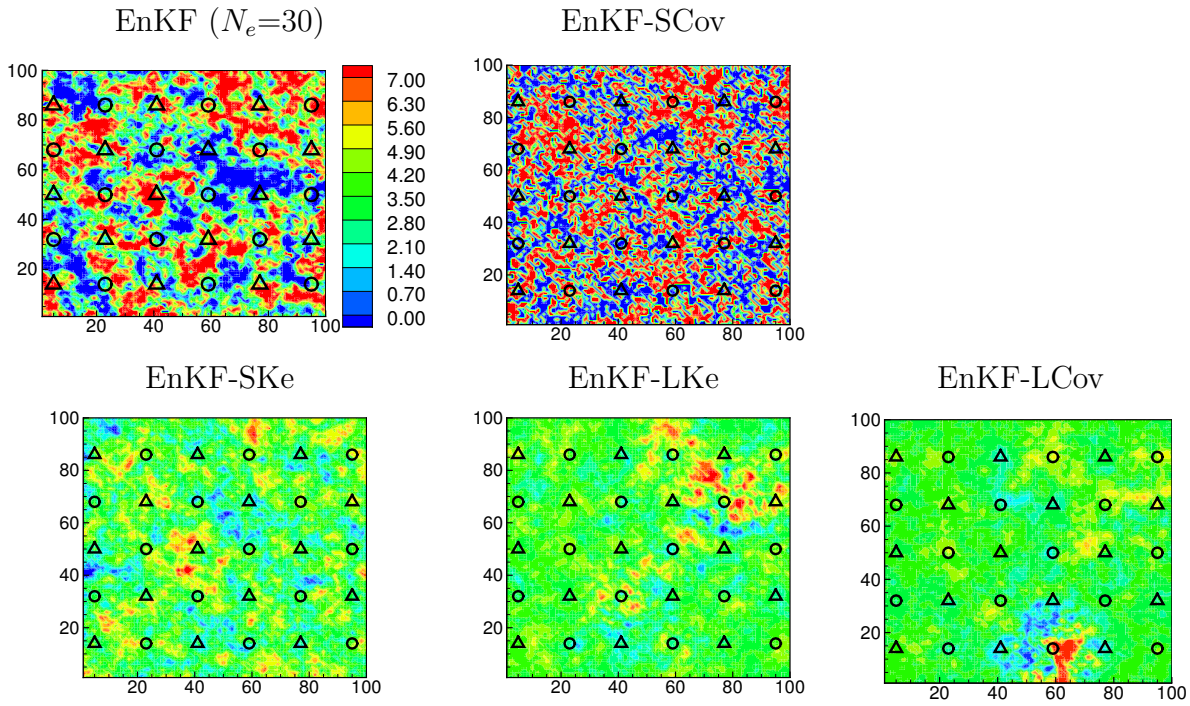


Figure 5.15: Final estimates of log permeability. (Mean over the ensemble after all data assimilation.)

Table 5.2: Statistical quantities of the final estimates of log permeability.

	Ensemble Mean	Ensemble STD	RMSE
EnKF (N_e=30)	3.53	0.005	2.98
EnKF-SCov	3.21	0.04	4.41
EnKF-SKe	3.50	0.15	1.25
EnKF-LKe	3.65	0.95	1.5
EnKF-LCov	3.53	0.72	1.24
EnKF (N_e=2000)	3.50	0.87	1.23

5.4.5 The estimates of Kalman gain

The Kalman gain contains the weighted correlations between data and variables in the state vector. For this 2D problem, there are 135 (45 data $\times$ 3 types of model variables)

different weighted correlation maps at each data assimilation step. Comparing each of them is time consuming. The effect of spurious correlations should be small in an ensemble of size 2000, which provides a good basis for comparison. Thus, the estimate of the Kalman gain from the EnKF with $N_e = 2000$ is treated as the true Kalman gain, and the RMSE of estimates from other methods are computed. The Kalman gain matrix is composed of 9 block matrices whose dimensions are 10,000 $\times$ 15 as shown in Eq. 5.19,

$$K_e = \left[\begin{array}{ccc|ccc|ccc} \ln K_1/\text{OPR 1} & \cdots & /\text{OPR 15} & /\text{WPR 1} & \cdots & /\text{WPR 15} & /\text{BHP 1} & \cdots & /\text{BHP 15} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ln K_{10000}/\text{OPR 1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hline P_1/\text{OPR 1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{10000}/\text{OPR 1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hline S_{w_1}/\text{OPR 1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{w_{10000}}/\text{OPR 1} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array}\right], \tag{5.19}$$

where $\ln K$ is log permeability, P is pressure, S_w is water saturation, OPR is the oil production rate of a producer, WPR is the water production rate of a producer, and BHP is the bottom hole pressure of an injector. An average RMSE of each block matrix with respect to the Kalman gain obtained from the standard EnKF with $N_e = 2000$ is calculated at two representative times to evaluate the quality of the Kalman gain estimate for a particular type of data and model variable.

For the early time, before water arriving at producers from injectors (mainly the first 2 data assimilation times), the estimates of C_{dd}^f are diagonally dominant, because the predicted data are not highly correlated with other predicted data. Table 5.3 shows the average RMSE of the Kalman gain estimates at data assimilation time 1.

Because the observed values of WPR are nearly zero, the average RMSE values are not calculated for the blocks related to either WPR or S_w. The standard EnKF has the highest average RMSE values for the estimates of the Kalman gain. The other methods provide comparatively good estimates of the Kalman gain.

Table 5.3: Average RMSE of the Kalman gain estimates at data assimilation time 1.

Variable/Data	EnKF (N_e=30)	EnKF-SCov	EnKF-SKe	EnKF-LKe	EnKF-LCov
$\ln K$/OPR	0.0020	0.0008	0.0008	0.0008	0.0007
$\ln K$/BHP	0.0004	0.0001	0.0001	0.0001	0.00006
P/OPR	0.65	0.57	0.50	0.53	0.57
P/BHP	0.1	0.05	0.03	0.03	0.02

By data assimilation time 4, most of the producers from the reference model show significant water production and the correlations between different data become stronger. Table 5.4 shows the average RMSE of the Kalman gain estimates at data assimilation time 4. In this table, the results from the standard EnKF (N_e=30) are not included, because the updated ensemble from the EnKF lost almost all the ensemble variability at the 1st data assimilation time and the estimates of the Kalman gain at later assimilation times do not contain any information. Comparing these four methods in Table 5.4, EnKF-SCov shows much larger RMSE values than those obtained from the other three methods, and EnKF-SKe generally results in the lowest RMSE values, especially for $\ln K$/WPR, P/WPR, and S_w/WPR.

5.4.6 Simultaneous estimation of spatially correlated and uncorrelated model parameters

In the previous reservoir data assimilation example, we estimated log permeability, which is a spatially correlated model parameter for which distance-dependent localization might be expected to work well. There are, however, sometimes model parameters to be estimated for which the concept of distance is not meaningful. In

Table 5.4: Average RMSE of the Kalman gain estimates at data assimilation time 4.

	EnKF-SCov	EnKF-SKe	EnKF-LKe	EnKF-LCov
$\ln K$/OPR	0.04	0.001	0.002	0.001
$\ln K$/WPR	5.4	0.01	0.2	1.1
$\ln K$/BHP	0.002	0.0002	0.0002	0.0002
P/OPR	12.14	0.67	0.72	0.69
P/WPR	1629.1	3.3	36.7	1726.9
P/BHP	0.58	0.08	0.07	0.09
S_w/OPR	0.0038	0.0002	0.0002	0.0002
S_w/WPR	0.5	0.001	0.01	0.09
S_w/BHP	0.0002	0.00002	0.00002	0.00002

this section, faults with unknown transmissibilities are incorporated into the reservoir model that was used in the previous example. The objective of this test is to see how the presence of spatially uncorrelated parameters in the state vector affects the localization and screening algorithms involved in the EnKF process. The EnKF-SCov method is not evaluated here because of its poor performance in the previous example.

All test settings are the same as those used in the previous example, except that 10 faults are incorporated in the reservoir model as shown in Fig. 5.16, and that the fault transmissibility multipliers for these 10 faults are to be estimated along with the log permeability at 10,000 gridblocks. The fault geometry is kept simple (rectangular) and all the gridblocks contained in one fault body are assumed to have the same fault transmissibility multiplier. The initial ensemble of the transmissibility multipliers of 10 faults is generated from a uniform distribution between 0.0 and 0.1.

At each data assimilation time, truncation was used to maintain the updated fault transmissiblity multiplier values within the range of 0.0 and 1.0. The final updated fault transmissibility multipliers obtained at the end of the data assimilation process are shown in Fig. 5.17. The ensemble estimates of fault transmissibility multipliers from the standard EnKF ($N_e = 30$) have collapsed to values that are quite far from

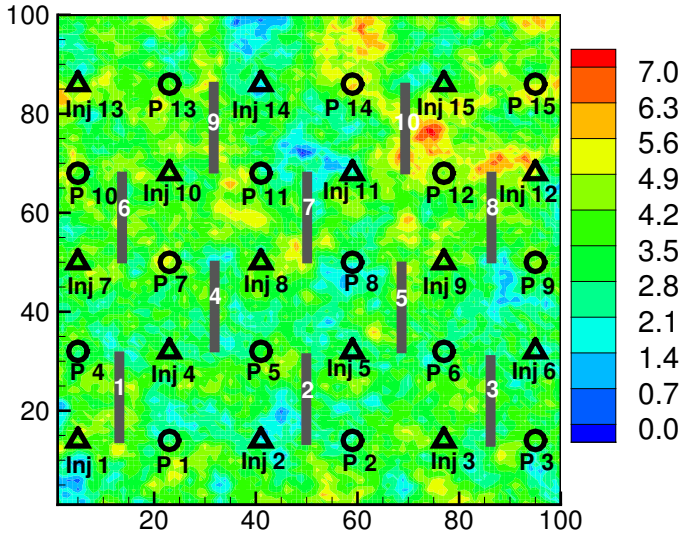


Figure 5.16: Reference log permeability field with 10 faults.

the reference values. The values of some estimates are substantially larger than 0.1. The estimates of fault transmissibility multipliers from EnKF-LKe are similar to those obtained from the standard EnKF ($N_e = 30$), because Kalman gain localization can not be applied on fault transmissibility multipliers. The EnKF with covariance localization (EnKF-LCov) shows the worst estimates of fault transmissibility multipliers. The localization applied on C_{dd}^f seems to have a negative influence on the updating of fault transmissibility multipliers. The EnKF-SKe method provides the best estimates of fault transmissibility multipliers, although the estimates of multipliers for faults 6 and 10 are poor.

The estimates of log permeability from EnKF-LKe and EnKF-LCov appear to have some extreme values (Fig. 5.18) and EnKF-LCov shows very strong artifacts of the localization. The localization coefficients for EnKF-LCov are based on the distance-dependence assumption which may be further weakened in the presence of flow barriers in the reservoir model. Table 5.5 shows that EnKF-LCov and EnKF-LKe result in larger RMSE values compared to the EnKF-SKe method. The ensemble STD from EnKF-SKe shows the same value as obtained from the previous example,

while the ensemble STD values from EnKF-LCov and EnKF-LKe vary slightly from their values obtained in the previous example.

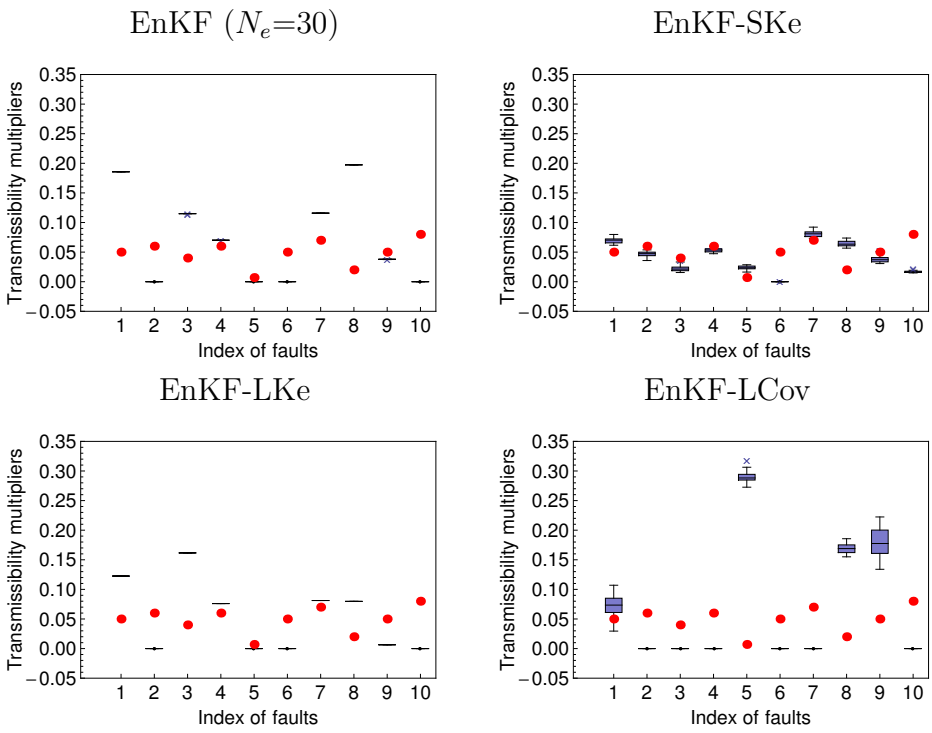


Figure 5.17: Final updated transmissibility multipliers (In the whisker box plot, red dots denote the true transmissibility multipliers).

Table 5.5: Statistical quantities of the final estimates of log permeability for the example with fault transmissibility multipliers.

	EnKF (N_e=30)	EnKF-SKe	EnKF-LKe	EnKF-LCov
Ensemble STD	0.006	0.15	0.97	0.70
RMSE	2.29	1.21	1.56	1.40

With the final updated log permeability fields and fault transmissibility multipliers, we rerun the simulations from time 0 to the end of the production period (day 520). Table 5.6 shows the error and spread of the data predictions from different methods. The EnKF-SKe results in the lowest $\hat{e}_d$ for OPR and WPR data, but shows slightly higher values for BHP data.

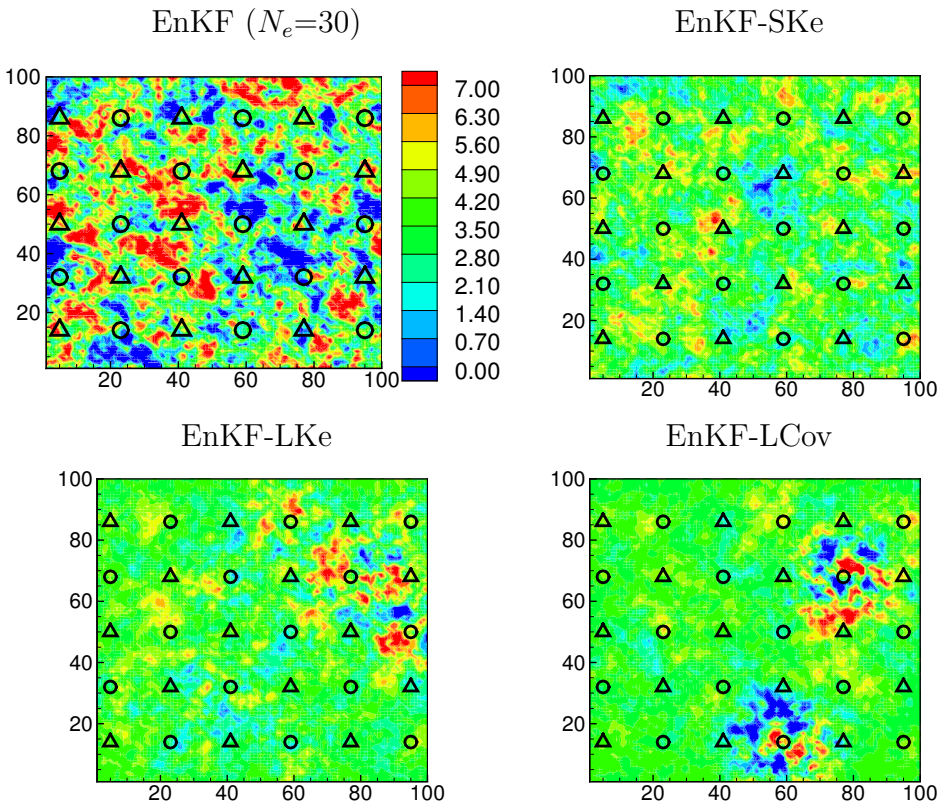


Figure 5.18: Final estimates of log permeability for the example with fault transmissibility multipliers.

Table 5.6: The average error versus average spread ($\hat{e}_d/\hat{\sigma}_d$) of the predictions of three types of data for the example with fault transmissibility multipliers.

	OPR (stb/day)	WPR (stb/day)	BHP (psi)
EnKF (N_e=30)	125/0.6	88/0.6	215/0.8
EnKF-SKe	64/19	41/15	213/46
EnKF-LKe	74/47	213/36	117/75
EnKF-LCov	77/35	106/26	82/51

Compared to the standard EnKF of the same ensemble size, the EnKF with screening or localization introduces extra computational cost for calculating the Kalman gain of bootstrapped ensembles and for computing the Schur product of the Kalman gain and screening/localizing factors. The extra computational cost of resampling, however, does not necessarily slow down the data assimilation process. On the contrary, Table 5.7 shows that EnKF with screening Kalman gain (EnKF-SKe) required only about 65% of the total CPU time needed for the standard EnKF without screening or localization. The reduction in computational time in the case of EnKF-SKe can be attributed to improved updates to model parameters leading to faster convergence of the Newton or linear iterations for solving the system equations inside the reservoir simulator.

Table 5.7: Total CPU time required for data assimilation and final rerun using 3 processors for the example with fault transmissibility multipliers.

	EnKF (N_e=30)	EnKF-SKe	EnKF-LKe	EnKF-LCov
CPU time (minutes)	24	16	15	12

5.5 Chapter summary

In this work, we evaluated and compared several methods of regularizing the Kalman gain and regularizing the covariance matrices used for computation of the Kalman gain. The performance of the methods was based on improvement in the estimates of the Kalman gain, quality of data prediction, and the estimates of model variables. Distance-dependent localization and bootstrap-based screening were both evaluated. The error analysis carried out in the 1D linear example showed that covariance regularization is more error prone than the Kalman gain regularization. This point is clearly illustrated by the dramatically different performances of Kalman gain screening and covariance screening. The performances of the distance-based covariance localization and Kalman gain localization, however, are not significantly different when

the state vector contains only spatially correlated variables. This is probably because knowledge of the correlation length, sensitivity, and well pattern used for constructing the localization coefficients substantially reduces the error in the coefficients.

We also showed that when regularization is applied to the covariance matrices, a consistency condition must be satisfied. For the problem of assimilating multiple, non-local observations, it is difficult to satisfy the consistency condition for the distance-dependent covariance localization. In the 2D nonlinear example, an approximately consistent form of covariance localization was applied with acceptable results in terms of matching data and maintaining ensemble effective rank. Some extreme values were observed in the final estimates of log permeability fields, however, especially for the case of estimating fault transmissibility multipliers. Certainly, these extreme values are not only caused by inconsistency, but also the assumption that the true correlations can be localized spatially. One key limitation of distance-based localization is that, when the distance from a gridblock to the data location is beyond the specified range, the correlation value at that gridblock is assumed to be zero, in which case, the Kalman gain value at that gridblock is determined only by the data whose correlation at that gridblock is non-zero. This can result in magnification of the influence from a particular data, which leads to over-correction on model variables. The distance-dependence assumption appears to be delicate and should probably be used with caution in the presence of complex geology.

The results from both 1D and 2D examples clearly show that screening Kalman gain (EnKF-SKe) worked well on a variety of problems with few assumptions. In the algorithm of screening Kalman gain, we directly calculate the replicate of Kalman gain from each bootstrap ensemble, and quantify the confidence level of the Kalman gain directly from the N_B bootstrap replicates of the Kalman gain. No assumption about the prior covariance is required in this case. The method can be used for estimating both spatially correlated and uncorrelated variables. Despite the apparent cost of

resampling the Kalman gain multiple times, the total computational cost for EnKF-SKe was less than that for standard EnKF because of reduced time for reservoir simulation.

CHAPTER VI

CONCLUSIONS

Sequential reservoir updating and performance prediction with valid uncertainty quantification continues to be a main direction of assisted history matching method, especially as the closed-loop reservoir management gets popular. In this direction, the ensemble Kalman filter (EnKF) is one of the most promising methods. This dissertation is focused on several challenging problems encountered in the practical application of the EnKF, which are related to the reduced-order representation of the covariance matrix, non-Gaussian prior distributions of model variables, and the implementation on large-scale oil field models. Although the work is for improving the performance of EnKF in reservoir engineering applications, the proposed methods and findings are also useful for other areas. In this chapter, the presented work is summarized.

In Chapter 2, a multiscale parameterization method for estimating non-Gaussian model parameters and better uncertainty quantification in reservoir characterization has been presented. With the multiscale parameterization, we showed that it is possible to update the multiscale features of reservoir properties using the ensemble Kalman filter by continual assimilation of production data. To avoid non-physical values of updated model variables, a transformation step is added in the framework of EnKF with multiscale parameterization. The applicability of the proposed method is verified by the successful history matching of the deepwater reservoir PFJ2. Without quantifying the uncertainty in regional trend, the water cut history of a main producer in PFJ2 was not be able to be matched, which evidently shows the importance of multiscale parameterization in real field applications.

The second major topic presented in the dissertation is the bootstrap-based screening algorithms that do not require pre-specifying localization range or evaluating large numbers of simulations. The investigation of the methods on covariance regularization and Kalman gain regularization resulted in three primary conclusions. First, if the localization of the two covariance matrices required for Kalman gain estimation are not consistent, the estimate of the Kalman gain will generally be poor at the observation location. The consistency condition can be difficult to apply for nonlocal observations. Second, the estimate of the Kalman gain that results from covariance regularization is generally subject to greater errors than the estimate of the Kalman gain that results from Kalman gain regularization. Third, in terms of removing spurious correlations in the estimation of spatially correlated variables, the performance of screening Kalman gain is comparable with the performance of localization methods (applied on either covariance or Kalman gain), but screening Kalman gain outperforms the distance-based localization methods in terms of generality for application, as the screening method can be used for estimating both spatially correlated and uncorrelated variables, and moreover, no assumption about the prior covariance is required during the process. The self-adaptive bootstrap-based screening methods seem promising as a default method for regularization in EnKF software, since it does not require any expertise on the part of the user. This is exactly the significance of this piece of work.

The performance of EnKF is being improved as new techniques are proposed, but several challenges still remain. First, the statistical error introduced by the limited ensemble size results in noise in the estimate of covariance and consequentially the estimate of Kalman gain. The distance-based localization methods and the proposed bootstrap-based screening methods can reduce the noise to a certain level, but definitely cannot eliminate the noise completely. Moreover, the absolute values of low-level true correlations, in the statistical sense, are likely to be underestimated

(or estimated with negative bias) using the screening methods, because the standard estimate of such low-level correlations based on a small ensemble is mostly off from the true correlation values. A statistical method that can detect the sign of the bias (positive or negative) in the absolute values of standard estimate might be useful, in which case, screening factor is only multiplied with the estimate having positive bias. In that situation, we might be able to improve the estimate of true weak correlations. Second, as data are assimilated, the loss of ensemble variability is usually greater than required by the data. In such scenario, the contributions of data assimilated earlier are larger than the recent data. Although methods have been proposed to partially solve this problem, including covariance inflation and using different Kalman gain to update different individual realization, more systematic investigations are required.

BIBLIOGRAPHY

Aanonsen, S. I., G. Nævdal, D. S. Oliver, A. C. Reynolds, and B. Vallès, Ensemble Kalman filter in reservoir engineering — a review, *SPE Journal*, **14**(3), 393–412, 2009.

Agbalaka, C. and D. S. Oliver, Application of the EnKF and localization to automatic history matching of facies distribution and production data, *Mathematical Geosciences*, **40**(4), 353–374, 2008.

Agbalaka, C. C. and D. S. Oliver, Automatic history matching of production and facies data with non-stationary proportions using EnKF, SPE 118916, in *Proceedings of the 2009 SPE Reservoir Simulation Symposium, The Woodlands, Texas*, 2009.

Anderson, J. L., Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, *Physica D: Nonlinear Phenomena*, **230**(1–2), 99–111, 2007.

Bergemann, K. and S. Reich, Localization techniques for ensemble transform Kalman filters, *submitted*, 2009.

Bianco, A., A. Cominelli, L. Dovera, G. Nævdal, and B. Vallès, History matching and production forecast uncertainty by means of the ensemble Kalman filter: A real field application (SPE-107161), in *SPE Europec/EAGE Annual Conference and Exhibition. Society of Petroleum Engineers. London, UK*, 2007.

Brouwer, D. R., G. Nævdal, J. D. Jansen, E. H. Vefring, and C. P. J. W. van Kruijsdijk, Improved reservoir management through optimal control and continuous model updating, in *SPE Annual Technical Conference and Exhibition, 26-29 September, Houston, Texas*, 2004.

Burgers, G., P. van Leeuwen, and G. Evensen, Analysis scheme in the ensemble Kalman filter, *Monthly Weather Review*, **126**(6), 1719–1724, 1998.

Capen, E. C., The difficulty of assessing uncertainty, *J. Petroleum Technology*, **28**(8), 843–850, 1976.

Chang, H., Y. Chen, and D. Zhang, Data assimilation of coupled fluid flow and geomechanics via ensemble Kalman filter (SPE 118963), *SPE Journal*, **accepted**, 2010.

Chen, Y. and D. S. Oliver, Cross-covariances and localization for EnKF in multiphase flow data assimilation, *Computational Geosciences*, **Online First**, 2009.

Chen, Y. and D. S. Oliver, Ensemble-based closed-loop optimization applied to Brugge Field (SPE 118926), *SPE Reservoir Evaluation & Engineering*, **13**(1), 56–71, 2010.

Chen, Y., D. S. Oliver, and D. Zhang, Data assimilation for nonlinear problems by ensemble Kalman filter with reparameterization, *Journal of Petroleum Science and Engineering*, **submitted**, 2007.

Chen, Y., D. S. Oliver, and D. Zhang, Efficient ensemble-based closed-loop production optimization, *SPE Journal*, **14**(4), 634–645, 2009.

Deutsch, C. V., *Geostatistical Reservoir Modeling*, first edn., Oxford, 2002.

Efron, B. and R. J. Tibshirani, *An Introduction to the Bootstrap*, first edn., Chapman & Hall/CRC, 1993.

Evensen, G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *Journal of Geophysical Research*, **99**(C5), 10,143–10,162, 1994.

Evensen, G., Sampling strategies and square root analysis schemes for the EnKF, *Ocean Dynamics*, **54**(6), 539–560, 2004.

Evensen, G., *Data Assimilation: The Ensemble Kalman Filter*, Springer, 2006.

Evensen, G., J. Hove, H. C. Meisingset, E. Reiso, K. S. Seim, and Ø. Espelid, Using the EnKF for assisted history matching of a North Sea reservoir model (SPE 106184), in *Proceedings of the 2007 SPE Reservoir Simulation Symposium,*, 2007.

Friedman, J. H., Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**(405), 165–175, 1989.

Furrer, R. and T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, *J. Multivar. Anal.*, **98**(2), 227–255, 2007.

Gao, G., M. Zafari, and A. C. Reynolds, Quantifying uncertainty for the PUNQ-S3 problem in a Bayesian setting with RML and EnKF, *SPE Journal*, **11**(4), 506–515, 2006.

Gaspari, G. and S. E. Cohn, Construction of correlation functions in two and three dimensions, *Quarterly Journal of the Royal Meteorological Society*, **125**(554), 723–757, 1999.

Gu, Y. and D. S. Oliver, History matching of the PUNQ-S3 reservoir model using the ensemble Kalman filter, *SPE Journal*, **10**(2), 51–65, 2005.

Gu, Y. and D. S. Oliver, An iterative ensemble Kalman filter for multiphase fluid flow data assimilation, *SPE Journal*, **12**(4), 438–446, 2007.

Hacker, J. P., J. L. Anderson, and M. Pagowski, Improved vertical covariance estimates for ensemble-filter assimilation of near-surface observations, *Monthly Weather Review*, **135**(3), 1021–1036, 2007.

Hamill, T. M., J. S. Whitaker, and C. Snyder, Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter, *Monthly Weather Review*, **129**(11), 2776–2790, 2001.

Haugen, V., G. Naevdal, L.-J. Natvik, G. Evensen, A. M. Berg, and K. M. Flornes, History matching using the ensemble Kalman filter on a North Sea field case, *SPE Journal*, **13**(4), 382–391, 2008.

Houtekamer, P. L. and H. L. Mitchell, Data assimilation using an ensemble Kalman filter technique, *Monthly Weather Review*, **126**(3), 796–811, 1998.

Houtekamer, P. L. and H. L. Mitchell, A sequential ensemble Kalman filter for atmospheric data assimilation, *Monthly Weather Review*, **129**(1), 123–137, 2001.

Houtekamer, P. L., H. L. Mitchell, G. Pellerin, M. Buehner, M. Charron, L. Spacek, and B. Hansen, Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations, *Monthly Weather Review*, **133**(3), 604–620, 2005.

Jafarpour, B. and D. B. McLaughlin, History matching with an ensemble Kalman filter and discrete cosine parameterization, *Computational Geosciences*, **12**(2), 227–244, 2008.

Kepert, J. D., Localisation, balance and choice of analysis variable in an ensemble Kalman filter, in *7th Adjoint Workshop*, vol. 12, 2006.

Lawniczak, W., R. Hanea, A. Heemink, and D. McLaughlin, Multiscale ensemble filtering for reservoir engineering applications, *Computational Geosciences*, **13**(2), 245–254, 2009.

Liu, N. and D. S. Oliver, Critical evaluation of the ensemble Kalman filter on history matching of geologic facies, *SPE Reservoir Evaluation & Engineering*, **8**(6), 470–477, 2005a.

Liu, N. and D. S. Oliver, Ensemble Kalman filter for automatic history matching of geologic facies, *Journal of Petroleum Science and Engineering*, **47**(3–4), 147–161, 2005b.

Lorenc, A. C., The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var, *Quarterly Journal of the Royal Meteorological Society*, **129**(595), 3183–3203, 2003.

Lorentzen, R. J., A. M. Berg, G. Nævdal, and E. H. Vefring, A new approach for dynamic optimization of waterflooding problems, in *SPE Intelligent Energy Conference and Exhibition*, 2006.

Lorentzen, R. J., K. K. Fjelde, J. Frøyen, A. C. V. M. Lage, G. Nævdal, and E. H. Vefring, Underbalanced drilling: Real time data interpretation and decision support, in *SPE/IADC Drilling Conference*, 2001.

Nævdal, G., D. R. Brouwer, and J.-D. Jansen, Waterflooding using closed-loop control, *Computational Geosciences*, **10**(1), 37–60, 2006.

Nævdal, G., L. M. Johnsen, S. I. Aanonsen, and E. H. Vefring, Reservoir monitoring and continuous model updating using ensemble Kalman filter, *SPE 84372*, 2003.

Nævdal, G., L. M. Johnsen, S. I. Aanonsen, and E. H. Vefring, Reservoir monitoring and continuous model updating using ensemble Kalman filter, *SPE Journal*, **10**(1), 66–74, 2005.

Nævdal, G., T. Mannseth, and E. H. Vefring, Near-well reservoir monitoring through ensemble Kalman filter: SPE 75235, in *Proceedings of SPE/DOE Improved Oil Recovery Symposium*, 2002.

Oliver, D. S. and Y. Chen, Recent progress on reservoir history matching: a review, *Computational Geosciences*, **submitted**, 2010.

Oliver, D. S., A. C. Reynolds, and N. Liu, *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, first edn., Cambridge University Press, Cambridge, 2008.

Oliver, D. S., Y. Zhang, H. Phale, and Y. Chen, Distributed parameter and state estimation in petroleum reservoirs, *Computers & Fluids*, **submitted**, 2010.

Omre, H., Bayesian Kriging — merging observations and qualified guesses in Kriging, *Mathematical Geology*, **19**(1), 25–39, 1987.

Sakov, P. and P. R. Oke, Implications of the form of the ensemble transformation in the ensemble square root filters, *Monthly Weather Review*, **136**(3), 1042–1053, 2008.

Schlumberger, ECLIPSE 100 Technical Description 2007A, GeoQuest., 2007.

Seiler, A., J. Rivenæs, S. Aanonsen, and G. Evensen, Structural uncertainty modelling and updating by production data integration, (SPE 125352), in *SPE/EAGE Reservoir Characterization and Simulation Conference, 19–21 October, Abu Dhabi, UAE*, 2009.

Skjervheim, J.-A., G. Evensen, S. I. Aanonsen, B. O. Ruud, and T. A. Johansen, Incorporating 4D seismic data in reservoir simulation models using ensemble Kalman filter, *SPE Journal*, **12**(3), 282–292, 2007.

Thulin, K., G. Li, S. I. Aanonsen, and A. C. Reynolds, Estimation of initial fluid contacts by assimilation of production data with EnKF, in *Proceedings of the 2007 SPE Annual Technical Conference and Exhibition*, 2007.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, Ensemble square-root filters, *Monthly Weather Review*, **131**, 1485–1490, 2003.

Vallès and G. Nævdal, Comparing different ensemble Kalman filter approaches, in *11th European Conference on the Mathematics of Oil Recovery*, EAGE, 2008.

Wang, C., G. Li, and A. C. Reynolds, Production optimization in closed-loop reservoir management (SPE-109805), in *SPE Annual Technical Conference and Exhibition*, 2007.

Whitaker, J. S. and T. M. Hamill, Ensemble data assimilation without perturbed observations, *Monthly Weather Review*, **130**(7), 1913–1924, 2002.

Yang, K., K. S. Tan, K. Kramer, and J. M. Yarus, Stochastic 3D geological modeling of deep water reservoir: Methodology and case studies, *Gulf Coast Association of Geological Societies Transactions*, **L**, 463–472, 2000.

Zafari, M., G. Li, and A. C. Reynolds, Iterative forms of the ensemble Kalman filter, in *Proceedings of the 10th European Conference on the Mathematics of Oil Recovery — Amsterdam*, p. A030, 2006.

Zafari, M. and A. C. Reynolds, Assessing the uncertainty in reservoir description and performance predictions with the ensemble Kalman filter, *SPE Journal*, **12**(3), 382–391, 2007.

Zhang, Y., N. Liu, and D. S. Oliver, Ensemble filter methods with perturbed observations applied to nonlinear problems, *Computational Geosciences*, **14**(2), 249–261, 2010.

Zhang, Y. and D. S. Oliver, History matching using a hierarchical stochastic model with the ensemble Kalman filter: A field case study, SPE-118879, in *Proceedings of the 2009 SPE Reservoir Simulation Symposium, The Woodlands, February 2–4*, 2009.

Zhang, Y. and D. S. Oliver, Evaluation and error analysis: Kalman gain regularization versus covariance regularization, *Computational Geosciences*, **submitted**, 2010a.

Zhang, Y. and D. S. Oliver, Improving the ensemble estimate of the Kalman gain by bootstrap sampling, *Mathematical Geosciences*, **42**(3), 327–345, 2010b.

APPENDIX A

DERIVATION OF SCREENING FACTOR

A.1 Point-wise estimate without regularization (α)

The screening factor in the bootstrap version of hierarchical filter is defined to minimize the squared difference of estimated Kalman gain matrices obtained from the N_B bootstrapped ensembles. The following derivation (similar as that is shown in Anderson (2007)) is for each entry in the Kalman gain, but the entry index (i, j) is neglected for convenience.

The screening factor α is defined to minimize,

$$\frac{1}{2}\sum_{q=1}^{N_B}\sum_{p=1,p\neq q}^{N_B}(\alpha K^*_{e_p} - K^*_{e_q})^2 \ , \tag{A-1}$$

where p and q are the indices of bootstrapped samples. The least square solution of Eq. A-1 is obtained by taking the 1st order derivative with respect to α and setting it equal to zero,

$$\alpha\sum_{q=1}^{N_B}\sum_{p=1,p\neq q}^{N_B}K^{*^2}_{e_p} - \sum_{q=1}^{N_B}\sum_{p=1,p\neq q}^{N_B}K^*_{e_p}K^*_{e_q} \ = \ 0 \ .$$

Solve the equation for α,

$$\begin{aligned}\alpha =& \frac{\sum_{q=1}^{N_B}\sum_{p=1,p\neq q}^{N_B}K^*_{e_p}K^*_{e_q}}{\sum_{q=1}^{N_B}\sum_{p=1,p\neq q}^{N_B}K^{*^2}_{e_p}} \\ =& \frac{(\sum_{p=1}^{N_B}K^*_{e_p})^2 - \sum_{p=1}^{N_B}K^{*^2}_{e_p}}{(N_B-1)\sum_{p=1}^{N_B}K^{*^2}_{e_p}} \\ =& \frac{1}{N_B-1}(\frac{(\sum_{p=1}^{N_B}K^*_{e_p})^2}{\sum_{p=1}^{N_B}K^{*^2}_{e_p}} - 1) \ .\end{aligned} \tag{A-2}$$

Using $\bar{K}_e$ to denote the mean of the N_B Kalman gain samples, we can represent $(\sum_{p=1}^{N_B} K_{e_p}^*)^2$ as

$$\begin{aligned}(\sum_{p=1}^{N_B} K_{e_p}^*)^2 =&(N_B\bar{K}_e)^2 \\ =&N_B^2\bar{K}_e^2 \ .\end{aligned} \tag{A-3}$$

The variance of the N_B Kalman gain samples is calculated as

$$\hat{\sigma}_k^2 = \frac{1}{N_B}\sum_{p=1}^{N_B}(K_{e_p}^* - \bar{K}_e)^2 \ ,$$

which can be rearranged as

$$\begin{aligned}N_B\hat{\sigma}_k^2 =&\sum_{p=1}^{N_B}(K_{e_p}^{*^2} - 2\bar{K}_eK_{e_p}^* + \bar{K}_e^2) \\ =&\sum_{p=1}^{N_B} K_{e_p}^2 - 2\bar{K}_e(\sum_{p=1}^{N_B} K_{e_p}) + \sum_{p=1}^{N_B}\bar{K}_e^2 \\ =&\sum_{p=1}^{N_B} K_{e_p}^{*^2} - 2N_B\bar{K}_e^2 + N_B\bar{K}_e^2 \\ =&\sum_{p=1}^{N_B} K_{e_p}^{*^2} - N_B\bar{K}_e^2 \ .\end{aligned} \tag{A-4}$$

Rearranging Eq. A-4, $\sum_{p=1}^{N_B} K_{e_p}^{*^2}$ can be expressed as,

$$\sum_{p=1}^{N_B} K_{e_p}^{*^2} = N_B(\hat{\sigma}_k^2 + \bar{K}_e^2) \ . \tag{A-5}$$

Substituting Eq. A-3 and Eq. A-5 into Eq. A-2, we obtain

$$\begin{aligned}\alpha \ &= \ \frac{1}{N_B - 1}(\frac{N_B^2\bar{K}_e^2}{N_B(\hat{\sigma}_k^2 + \bar{K}_e^2)} - 1) \\ &= \ \frac{1}{N_B - 1}(\frac{N_B}{\hat{\sigma}_k^2/\bar{K}_e^2 + 1} - 1) \\ &= \ \frac{1}{N_B - 1}(\frac{N_B}{\hat{C}_v^2 + 1} - 1) \\ &= \ \frac{1 - \hat{C}_v^2/(N_B - 1)}{1 + \hat{C}_v^2} \ .\end{aligned}$$

A.2 Regularized point-wise estimate (α_r)

The objective function to be minimized is

$$\begin{aligned} S(\alpha_r) &= \frac{1}{2N_B}\sum_{p=1}^{N_B} \parallel (\alpha_r \circ K^*_{e_p} - \bar{K}_e) \circ \lambda_k \parallel_F^2 + \frac{1}{2} \parallel \alpha_r \circ \lambda_\alpha \parallel_F^2 \\ &= \frac{1}{2}(\sum_i \sum_j \frac{\sum_{p=1}^{N_B}(\alpha_{r_{i,j}} K^{*^p}_{e_{i,j}} - \bar{K}_{e_{i,j}})^2}{N_B \hat{\sigma}^2_{k_{i,j}}}) + \frac{1}{2}(\sum_i \sum_j \frac{\alpha^2_{r_{i,j}}}{\sigma^2_\alpha}) \ . \end{aligned} \tag{A-6}$$

Minimizing Eq. A-6 is equivalent to minimize the following objective function for individual entry in the Kalman gain matrix (where the subscrits i and j are ignored for convenience),

$$S(\alpha_r) = \frac{\sum_{p=1}^{N_B}(\alpha_r K^*_{e_p} - \bar{K}_e)^2}{2N_B \hat{\sigma}^2_k} + \frac{\alpha^2_r}{2\sigma^2_\alpha} \ ,$$

Taking the 1st order derivative of $S(\alpha_r)$ with respect to α_r and setting it to zero, we obtain

$$(\frac{\sum_{p=1}^{N_B} K^{*^2}_{e_p}}{N_B \hat{\sigma}^2_k} + \frac{1}{\sigma^2_\alpha})\alpha_r - \frac{\bar{K}_e \sum_{p=1}^{N_B} K^*_{e_p}}{N_B \hat{\sigma}^2_k} = 0 \ . \tag{A-7}$$

Substituting Eq. A-5 and $\sum_{p=1}^{N_B} K^*_{e_p} = N_B \bar{K}_e$ into Eq. A-7 and rearranging, we can get the final form for α_r,

$$\alpha_r = \frac{1}{1 + (1 + 1/\sigma^2_\alpha)\hat{C}^2_v} \ .$$

A.3 Smooth estimate (α_s)

The objective function is defined as

$$S(\alpha_s) = \frac{1}{2N_B}\sum_{p=1}^{N_B} \parallel (\alpha_s \circ K^*_{e_p} - K_e) \circ \lambda_k \parallel_F^2 + \frac{1}{2}(\alpha_s^T (W^T W + \frac{1}{\sigma^2_\alpha} I)\alpha_s) \ .$$

The optimal condition is

$$(I + \Gamma)\alpha_s - \gamma + (W^T W + \frac{1}{\sigma^2_\alpha} I)\alpha_s = 0 \ , \tag{A-8}$$

where Γ is a diagonal matrix with diagonal element $\Gamma_{i,i} = \frac{1}{\hat{C}^{2v_i}}$, and $\gamma = [\frac{1}{\hat{C}^2_{v_1}}, \frac{1}{\hat{C}^2_{v_2}}, \cdots, \frac{1}{\hat{C}^2_{v_{Nm}}}]^T$. Rearranging Eq. A-8 leads to

$$(W^T W + \Lambda)\alpha_s = \gamma \ ,$$

where Λ is a diagonal matrix with each element $\Lambda_{i,i} = \frac{1}{\sigma^2_\alpha} + 1 + \frac{1}{\hat{C}^2_{v_i}}$. We can also express α_s as

$$\alpha_s = (W^T W + \Lambda)^{-1}\gamma \ .$$